Introductory lecture: examples of optimization problems, optimization methods

Optimization methods in machine learning

Aleksandr Beznosikov

Innopolis University

4 September 2023

Lecturer

- Aleksandr Nikolaevich Beznosikov
- PhD at MIPT
- Lecturer on optimization at MIPT
- 10+ papers about optimization at conference and in journals on machine learning
- lab head at MIPT, researcher at Innopolis, Skoltech, HSE, Yandex, ITTP RAS, MBZUAI (UAE)
- website: anbeznosikov.github.io
- emails: beznosikov.an@phystech.edu, anbeznosikov@gmail.com
- tg: @abeznosikov

Grading

- Homeworks (70 %). 3 assignments for 2-3 weeks each. Deadlines are strict.
- Project (30 %). There are two types of projects in the course: 1) read 1-3 papers and implement methods from there, test their workability (check project progress every two weeks via github commits), 2) real project aimed at writing a scientific paper (call and discussion every week).
- A $\geq$ 80%, B $\geq$ 60 %, C $\geq$ 40%.

Recommended literature on the topic of the lecture

This lecture is mostly based on the book by Y.E. Nesterov

- Нестеров, Юрий Евгеньевич. «Введение в выпуклую оптимизацию» (2010). URL: `https://mipt.ru/dcam/upload/abb/nesterovfinal-arpgzk47dcy.pdf`
- Yurii Nesterov. «Introductory Lectures on Convex Optimization: A Basic Course»

ML approach

Question: How is the optimization problem in machine learning formulated?

ML approach

Question: How is the optimization problem in machine learning formulated?

- In practice, only a finite sample/final data set $\xi_1, \ldots, \xi_m$ is usually available, on which the problem of minimizing the empirical loss function is formulated:

$$\min_{x \in Q \subseteq \mathbb{R}^d} \left\{ \hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} f(x, \xi_i) = \frac{1}{m} \sum_{i=1}^{m} \ell(g(x, \xi_{i,data}), \xi_{i,label}) \right\}. \tag{1}$$

ML approach

Question: How is the optimization problem in machine learning formulated?

- In practice, only a finite sample/final data set $\xi_1, \ldots, \xi_m$ is usually available, on which the problem of minimizing the empirical loss function is formulated:

$$\min_{x \in Q \subseteq \mathbb{R}^d} \left\{ \hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} f(x, \xi_i) = \frac{1}{m} \sum_{i=1}^{m} \ell(g(x, \xi_{i,data}), \xi_{i,label}) \right\}. \tag{1}$$

- However, it is assumed that the sample came from some distribution $\mathcal{D}$ and one wants to get a good approximate solution to the problem of minimizing the expected loss function:

$$\min_{x \in Q \subseteq \mathbb{R}^d} \left\{ f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)] = \mathbb{E}_{\xi \sim \mathcal{D}}[\ell(g(x, \xi_{data}), \xi_{label})] \right\}. \tag{2}$$

ML approach

- (1) is called offline formulation, (2) is typically called online formulation. **Question:** why?

ML approach

- (1) is called offline formulation, (2) is typically called online formulation. **Question:** why?
- In (1), we have data samples that can be treated any way we want, forgetting about the original distribution $\mathcal{D}$. In (2), there may be no samples. The data come online one point at a time. We need to process this data at the moment of arrival, there is no time to collect a large data set.

ML approach

- (1) is called offline formulation, (2) is typically called online formulation. **Question:** why?
- In (1), we have data samples that can be treated any way we want, forgetting about the original distribution $\mathcal{D}$. In (2), there may be no samples. The data come online one point at a time. We need to process this data at the moment of arrival, there is no time to collect a large data set. **Question:** A natural question arises – how are these problems related?
- (1) is a Monte Carlo approximation of the integral (2). (2) is more general than (1).

ML approach

- (1) is called offline formulation, (2) is typically called online formulation. **Question:** why?
- In (1), we have data samples that can be treated any way we want, forgetting about the original distribution $\mathcal{D}$. In (2), there may be no samples. The data come online one point at a time. We need to process this data at the moment of arrival, there is no time to collect a large data set. **Question:** A natural question arises – how are these problems related?
- (1) is a Monte Carlo approximation of the integral (2). (2) is more general than (1).

ML approach

The classic results from

Shalev-Shwartz, S., Shamir, O., Srebro, N. and Sridharan, K., 2009, June. Stochastic Convex Optimization. In COLT.
https://ttic.uchicago.edu/~nati/Publications/nonlinearTR.pdf

Feldman, V. and Vondrak, J., 2019, June. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In Conference on Learning Theory (pp. 1270-1279). PMLR.
http://proceedings.mlr.press/v99/feldman19a/feldman19a.pdf

give: if the functions $f(x, \xi)$ are convex and M-Lipschitz, Q has diameter D and $\hat{x}^* = \text{argmin}_{x \in Q} \hat{f}(x)$, then with probability at least $1 - \delta$

$$f(\hat{x}^*) - \min_{x \in Q} f(x) = O\left(\sqrt{\frac{M^2 D^2 n \ln(m) \ln(n/\delta)}{m}}\right).$$

Statistical approach: linear regression

Suppose that some variable y depends on the variables $a_2, a_3, \ldots, a_n$ in a linear manner:

$$y(a_2, \ldots, a_n) = x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n,$$

where the coefficients $x_1, \ldots, x_n$ are unknown to us. Suppose we want to find these coefficients by measuring the variable y at different values of $a_2, \ldots, a_n$. It would seem to be a problem, because it is enough to make n measurements to solve the system. **Question:** what problem?

Statistical approach: linear regression

Suppose that some variable y depends on the variables $a_2, a_3, \ldots, a_n$ in a linear manner:

$$y(a_2, \ldots, a_n) = x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n,$$

where the coefficients $x_1, \ldots, x_n$ are unknown to us. Suppose we want to find these coefficients by measuring the variable y at different values of $a_2, \ldots, a_n$. It would seem to be a problem, because it is enough to make n measurements to solve the system. **Question:** what problem?In reality, the measurements are made with some error: for a given set $a_2^i, a_3^i, \ldots, a_n^i$ we measure

$$y_i = x_1 + a_2^i x_2 + a_3^i x_3 + \ldots + a_n^i x_n + \xi_i,$$

where $\xi_i \sim \mathcal{N}(0, \sigma^2)$.

Statistical approach: linear regression

- In other words, we assume that $y_i \sim \mathcal{N}(x_1 + a_2^i x_2 + a_3^i x_3 + \ldots + a_n^i x_n, \sigma^2)$, where the parameters $x = (x_1, \ldots, x_n)^\top$ are required to be found from a finite simple sample $\{y_i\}_{i=1}^m$ (we will assume that $y_1, \ldots, y_n$ are independent random variables). **Question:** How better to choose the parameters $x_1, \ldots, x_n$? And what does better mean?

Statistical approach: linear regression

- In other words, we assume that $y_i \sim \mathcal{N}(x_1 + a^i_2 x_2 + a^i_3 x_3 + \ldots + a^i_n x_n, \sigma^2)$, where the parameters $x = (x_1, \ldots, x_n)^\top$ are required to be found from a finite simple sample $\{y_i\}_{i=1}^m$ (we will assume that $y_1, \ldots, y_n$ are independent random variables). **Question:** How better to choose the parameters $x_1, \ldots, x_n$? And what does better mean?
- We can, for example, consider the maximum likelihood estimator (for brevity we introduce the vector $a^i = (1, a^i_2, \ldots, a^i_n)^\top$):

$$
\begin{aligned}
\hat{x} &= \arg\max_{x \in \mathbb{R}^n} \prod_{i=1}^m \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \langle a^i, x\rangle)^2\right) \\
&= \arg\max_{x \in \mathbb{R}^n} \ln\left(\prod_{i=1}^m \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \langle a^i, x\rangle)^2\right)\right).
\end{aligned}
$$

Statistical approach: linear regression

- Since the logarithm of the product is the sum of logarithms and since additive and multiplicative constants do not change the point of minimum, we obtain:

$$\begin{aligned}\hat{x} &= \arg\max_{x\in\mathbb{R}^n}\left\{\text{Const} + \sum_{i=1}^{m} -\frac{1}{2\sigma^2}(y_i - \langle a^i, x\rangle)^2\right\}\\ &= \arg\min_{x\in\mathbb{R}^n}\sum_{i=1}^{m}(y_i - \langle a^i, x\rangle)^2\\ &= \arg\min_{x\in\mathbb{R}^n}\|\mathrm{A}x - y\|_2^2,\end{aligned}$$

where the matrix A is composed of rows $(a^i)^\top$.

Optimization problem

$$\min_{\substack{g_i(x)\&0,\\ i=1,\dots,m,\\ x\in Q}} f(x) \tag{3}$$

- $Q \subseteq \mathbb{R}^d$ — subset of d-dimensional space
- $f : Q \to \mathbb{R}$ — some function defined on the set Q
- Either $\leq$ or $=$ is taken as &
- $g_i(x) : Q \to \mathbb{R}$, $i = 1, \dots, m$ — constraint functions

Optimization tasks. First observations.

1. In general, optimization problems may not have a solution. For example, the problem $\min_{x \in \mathbb{R}} x$ has no solution.
2. Optimization problems often cannot be solved analytically.
3. Their complexity depends on the type of the target function f, the set Q and may depend on the dimensionality x.

Optimization tasks. First observations.

1. In general, optimization problems may not have a solution. For example, the problem $\min_{x \in \mathbb{R}} x$ has no solution.
2. Optimization problems often cannot be solved analytically.
3. Their complexity depends on the type of the target function f, the set Q and may depend on the dimensionality x.

If an optimization problem has a solution, then in practice it is usually solved, generally speaking, approximately. For this purpose, special algorithms are used, which are called optimization methods.

Optimization methods I

- There is no point in looking for the best method to solve a particular problem. For example, the best method for solving the problem $\min_{x \in \mathbb{R}^d} \|x\|^2$ converges for 1 iteration: this method simply always gives the answer $x^* = 0$. Obviously, this method is not suitable for other problems.
- The efficiency of a method is determined for a class of problems, since numerical methods are usually developed for the *approximate* solution of a set of similar problems.
- The method is developed for a class of problems $\Longrightarrow$ method may not have complete information about the problem from the beginning. Instead, the method uses a model of the problem, e.g., the problem formulation, a description of the functional components, the set on which the optimization takes place, etc.

Optimization methods II

- It is assumed that the numerical method can accumulate specific information about the problem by means of some *oracle*. The oracle can be understood as some device that answers successive questions of the numerical method.

Examples of oracles

- **Zeroth order oracle** at the requested point x returns the value of the target function $f(x)$.
- **First order oracle** at the requested point returns the value of the function $f(x)$ and its gradient at that point $\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right)$.

General iterative scheme of the optimization method $\mathcal{M}$

Input: initial point x^0 (0 – upper index), required accuracy of the problem solution $\varepsilon > 0$.

General iterative scheme of the optimization method $\mathcal{M}$

Input: initial point x^0 (0 – upper index), required accuracy of the problem solution $\varepsilon > 0$.
Settings. Set $k = 0$ (iteration counter) and $I_{-1} = \varnothing$ (accumulated information model of the problem to be solved).

General iterative scheme of the optimization method $\mathcal{M}$

Input: initial point x^0 (0 – upper index), required accuracy of the problem solution $\varepsilon > 0$.
Settings. Set $k = 0$ (iteration counter) and $I_{-1} = \varnothing$ (accumulated information model of the problem to be solved).
Main loop

1. Ask a question to the oracle $\mathcal{O}$ at the point x^k.

General iterative scheme of the optimization method $\mathcal{M}$

Input: initial point x^0 (0 – upper index), required accuracy of the problem solution $\varepsilon > 0$.
Settings. Set $k = 0$ (iteration counter) and $I_{-1} = \varnothing$ (accumulated information model of the problem to be solved).
Main loop

1. Ask a question to the oracle $\mathcal{O}$ at the point x^k.
2. Recalculate the information model: $I_k = I_{k-1} \cup (x_k, \mathcal{O}(x^k))$.

General iterative scheme of the optimization method $\mathcal{M}$

Input: initial point x^0 (0 – upper index), required accuracy of the problem solution $\varepsilon > 0$.
Settings. Set $k = 0$ (iteration counter) and $I_{-1} = \varnothing$ (accumulated information model of the problem to be solved).
Main loop

1. Ask a question to the oracle $\mathcal{O}$ at the point x^k.
2. Recalculate the information model: $I_k = I_{k-1} \cup (x_k, \mathcal{O}(x^k))$.
3. Apply the rule of the method $\mathcal{M}$ to obtain a new point x^{k+1} using the I_k model.

General iterative scheme of the optimization method $\mathcal{M}$

Input: initial point x^0 (0 – upper index), required accuracy of the problem solution $\varepsilon > 0$.
Settings. Set $k = 0$ (iteration counter) and $I_{-1} = \varnothing$ (accumulated information model of the problem to be solved).
Main loop

1. Ask a question to the oracle $\mathcal{O}$ at the point x^k.
2. Recalculate the information model: $I_k = I_{k-1} \cup (x_k, \mathcal{O}(x^k))$.
3. Apply the rule of the method $\mathcal{M}$ to obtain a new point x^{k+1} using the I_k model.
4. Check the stopping criterion $\mathcal{T}_\varepsilon$. If the criterion is met, output the answer $\bar{x}$, otherwise put $k := k + 1$ and get back to the step 1.

Examples of iterative methods. Gradient descent

Let us consider the optimization problem

$$\min_{x \in \mathbb{R}^d} f(x), \tag{4}$$

where the function $f(x)$ is differentiable. Suppose that at any point we can calculate its gradient.

Examples of iterative methods. Gradient descent

Let us consider the optimization problem

$$\min_{x \in \mathbb{R}^d} f(x), \tag{4}$$

where the function $f(x)$ is differentiable. Suppose that at any point we can calculate its gradient.

Алгоритм 1 Gradient descent with constant step size

Input: step size $\gamma > 0$, initial point $x^0 \in \mathbb{R}^d$, number of iterations N

1: **for** $k = 0, 1, \ldots, N-1$ **do**

2: Compute $\nabla f(x^k)$

3: $x^{k+1} = x^k - \gamma \nabla f(x^k)$

4: **end for**

Output: x^N

Question: what are an oracle $\mathcal{O}$, a model I and a stopping criterion here?

Examples of iterative methods. Newton's method

Let us consider the optimization problem

$$\min_{x \in \mathbb{R}^d} f(x), \tag{5}$$

where the function $f(x)$ is twice continuously differentiable. Suppose that at any point we can calculate its gradient and the matrix of second derivatives of $\nabla^2 f(x)$.

Examples of iterative methods. Newton's method

Let us consider the optimization problem

$$\min_{x \in \mathbb{R}^d} f(x), \tag{5}$$

where the function $f(x)$ is twice continuously differentiable. Suppose that at any point we can calculate its gradient and the matrix of second derivatives of $\nabla^2 f(x)$.

Алгоритм 2 Newton's method

Input: initial point $x^0 \in \mathbb{R}^d$, number of iterations N

1: **for** $k = 0, 1, \ldots, N-1$ **do**
2: Compute $\nabla f(x^k)$ и $\nabla^2 f(x^k)$
3: $x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$
4: **end for**

Output: x^N

Complexity of optimization methods

- **Analytic Complexity** — the number of oracle calls required to solve the problem with ε accuracy.
- **Arithmetic Complexity** — the total number of calculations (including oracle work) required to solve the problem with ε accuracy.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Consider the problem:

$$\min_{x \in B_d} f(x) \tag{6}$$

- $B_d = \{x \in \mathbb{R}^d \mid 0 \leq x_i \leq 1, \quad i = 1, \ldots, d\}$
- The function $f(x)$ is M-Lipschitz on B_d with respect to the ℓ_∞-norm:

$$\forall x, y \ |f(x) - f(y)| \leq M\|x - y\|_\infty = M \max_{i=1,\ldots,d} |x_i - y_i|. \tag{7}$$

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Observation

The set B_d is bounded and closed, i.e., a compact, and the Lipschitzness of the function f implies its continuity, the the problem (6) has a solution, because the function continuous on the compact reaches its minimum and maximum values. Let $f_* = \min_{x \in B_d} f(x)$.

- **Class of methods.** For this problem, we consider zero-order methods.
- **Goal.** Find $\bar{x} \in B_d$: $f(\bar{x}) - f_* \leq \varepsilon$.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Let us consider one of the simplest ways of solving this problem — the method of uniform enumeration.

Алгоритм 3 Uniform search method

Input: целочисленный параметр перебора $p \geq 1$

1: Сформировать $(p+1)^d$ точек вида $x_{(i_1,\ldots,i_d)} = \left(\frac{i_1}{p}, \frac{i_2}{p}, \ldots, \frac{i_d}{p}\right)^\top$, где $(i_1, \ldots, i_d) \in \{0, 1, \ldots, p\}^n$

2: Среди точек $x_{(i_1,\ldots,i_d)}$ найти точку $\bar{x}$ с наименьшим значением целевой функции f.

Output: $\bar{x}, f(\bar{x})$

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Theorem 1 (Theorem 1.1.1 from Nesterov's book)

The algorithm 3 with parameter p returns a point $\bar{x}$ such that

$$f(\bar{x}) - f_* \leq \frac{M}{2p}, \tag{8}$$

whence it follows that the uniform search method needs in the worst case

$$\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2 \right)^d \tag{9}$$

calls of the oracle to guarantee $f(\bar{x}) - f_* \leq \varepsilon$.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Proof of Theorem 1

Let x_* — the solution of the problem (the point of minimum of the function f). Then in the constructed «grid» of points there is a point $x_{(i_1,\ldots,i_d)}$ such that $x := x_{(i_1,\ldots,i_d)} \leq x^* \leq x_{(i_1+1,\ldots,i_d+1)} =: y$, where the «$\leq$» sign is applied component by component.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Proof of Theorem 1

Let x_* — the solution of the problem (the point of minimum of the function f). Then in the constructed «grid» of points there is a point $x_{(i_1,\ldots,i_d)}$ such that $x := x_{(i_1,\ldots,i_d)} \leq x^* \leq x_{(i_1+1,\ldots,i_d+1)} =: y$, where the «$\leq$» sign is applied component by component. First, $y_i - x_i = \frac{1}{p}$ and $x_i^* \in [x_i, y_i]$ for all $i = 1, \ldots, d$.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Proof of Theorem 1

Let x_* — the solution of the problem (the point of minimum of the function f). Then in the constructed «grid» of points there is a point $x_{(i_1,\ldots,i_d)}$ such that $x := x_{(i_1,\ldots,i_d)} \leq x^* \leq x_{(i_1+1,\ldots,i_d+1)} =: y$, where the «$\leq$» sign is applied component by component. First, $y_i - x_i = \frac{1}{p}$ and $x_i^* \in [x_i, y_i]$ for all $i = 1, \ldots, d$. Furthermore, consider the points $\hat{x}$ and $\tilde{x}$ such that $\hat{x} = \frac{x+y}{2}$ and

$$\tilde{x}_i = \begin{cases} y_i, & \text{if } x_i^* \geq \hat{x}_i, \\ x_i, & \text{otherwise.} \end{cases}$$

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Proof of Theorem 1 (continued)

Note that $\tilde{x}$ belongs to «grid» and $|\tilde{x}_i - x_i^*| \leq \frac{1}{2p}$, and hence $\|\tilde{x} - x^*\|_\infty \leq \frac{1}{2p}$.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Proof of Theorem 1 (continued)

Note that $\tilde{x}$ belongs to «grid» and $|\tilde{x}_i - x_i^*| \leq \frac{1}{2p}$, and hence $\|\tilde{x} - x^*\|_\infty \leq \frac{1}{2p}$. $f(\bar{x}) \leq f(\tilde{x})$ (by definition), we get

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Proof of Theorem 1 (continued)

Note that $\tilde{x}$ belongs to «grid» and $|\tilde{x}_i - x_i^*| \leq \frac{1}{2p}$, and hence $\|\tilde{x} - x^*\|_\infty \leq \frac{1}{2p}$. $f(\bar{x}) \leq f(\tilde{x})$ (by definition), we get

$$f(\bar{x}) - f_* \leq f(\tilde{x}) - f_* \leq M\|\tilde{x} - x^*\|_\infty \leq \frac{M}{2p}.$$

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Proof of Theorem 1 (continued)

Note that $\tilde{x}$ belongs to «grid» and $|\tilde{x}_i - x_i^*| \leq \frac{1}{2p}$, and hence $\|\tilde{x} - x^*\|_\infty \leq \frac{1}{2p}$. $f(\bar{x}) \leq f(\tilde{x})$ (by definition), we get

$$f(\bar{x}) - f_* \leq f(\tilde{x}) - f_* \leq M\|\tilde{x} - x^*\|_\infty \leq \frac{M}{2p}.$$

The above estimate is achieved by a uniform enumeration over $(p+1)^d$ calls to the oracle. Hence, to guarantee $f(\bar{x}) - f_* \leq \varepsilon$, we need to take $p = \left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 1$, i.e. the method will make $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2 \right)^d$ calls to the oracle.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

- Suppose $M = 2$, $d = 13$ and $\varepsilon = 0.01$, that is, the dimensionality of the problem is relatively small and the accuracy of solving the problem is not too high.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

- Suppose $M = 2$, $d = 13$ and $\varepsilon = 0.01$, that is, the dimensionality of the problem is relatively small and the accuracy of solving the problem is not too high.
- The required number of calls to the oracle:
$\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d = 102^{13} > 10^{26}$.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

- Suppose $M = 2$, $d = 13$ and $\varepsilon = 0.01$, that is, the dimensionality of the problem is relatively small and the accuracy of solving the problem is not too high.
- The required number of calls to the oracle: $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d = 102^{13} > 10^{26}$.
- Computational power: 10^{11} arithmetic operations per second.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

- Suppose $M = 2$, $d = 13$ and $\varepsilon = 0.01$, that is, the dimensionality of the problem is relatively small and the accuracy of solving the problem is not too high.
- The required number of calls to the oracle: $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d = 102^{13} > 10^{26}$.
- Computational power: 10^{11} arithmetic operations per second.
- Total time: at least 10^{15} seconds, which is over 30 million years.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

- **Question:** What did we get in Theorem 1? Is it a lower or an upper bound? What are lower or an upper bounds?
- The upper bounds are guarantees of the performance of a particular algorithm on a class of problems. In Theorem 1, we obtained upper bounds for our uniform search algorithm for Lipschitz functions.
- Lower bounds are the opposite of that. When obtaining lower bounds, we pick a particular bad function from the class and show that any algorithm on that function performs no better than the lower bound.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

- **Question:** is it possible to propose another method from the class under consideration, which will find an approximate solution much faster? Perhaps we have just proposed a bad method?

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

- **Question:** is it possible to propose another method from the class under consideration, which will find an approximate solution much faster? Perhaps we have just proposed a bad method? It turns out that it is not: nothing fundamentally better can be proposed.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

- **Question:** is it possible to propose another method from the class under consideration, which will find an approximate solution much faster? Perhaps we have just proposed a bad method? It turns out that it is not: nothing fundamentally better can be proposed.

Theorem 2 (Theorem 1.1.2 from Nesterov's book)

Let $\varepsilon < \frac{M}{2}$. Then the analytical complexity of the described class of problems, i.e., the analytical complexity of the method on the «worst» problem from this class is at least

$$\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor\right)^d \quad \text{oracle calls.} \tag{10}$$

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Scheme of the proof of Theorem 2

Let $p = \left\lfloor \frac{M}{2\varepsilon} \right\rfloor$. Prove from the contrary: suppose there exists a method that solves the problem in $N < p^d$ calls to the oracle to solve the problem with ε accuracy (by function).

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Scheme of the proof of Theorem 2

Let $p = \left\lfloor \frac{M}{2\varepsilon} \right\rfloor$. Prove from the contrary: suppose there exists a method that solves the problem in $N < p^d$ calls to the oracle to solve the problem with ε accuracy (by function). Let's construct such a function on which the method cannot find a ε-solution, with the help of an adversarial oracle: for every query of the method about the value of the function at the requested point, the oracle returns 0.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Scheme of the proof of Theorem 2

Let $p = \left\lfloor \frac{M}{2\varepsilon} \right\rfloor$. Prove from the contrary: suppose there exists a method that solves the problem in $N < p^d$ calls to the oracle to solve the problem with ε accuracy (by function). Let's construct such a function on which the method cannot find a ε-solution, with the help of an adversarial oracle: for every query of the method about the value of the function at the requested point, the oracle returns 0. Then by Dirichlet's principle there is such a «cube» $B = \{x \mid \hat{x} \leq x \leq \hat{x} + \frac{1}{p}e\}$, где $\hat{x}$ и $\hat{x} + \frac{1}{p}e$ — points from «grid» with step p, e — unit vector. **Question:** what is Dirichlet's principle?

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Scheme of the proof of Theorem 2 (continued)

Let x^* — be the center of the «cube» B, i.e., $x^* = \hat{x} + \frac{1}{2p}e$.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Scheme of the proof of Theorem 2 (continued)

Let x^* — be the center of the «cube» B, i.e., $x^* = \hat{x} + \frac{1}{2p}e$. Consider the function $\bar{f}(x) = \min\{0, M\|x - x^*\|_\infty - \varepsilon\}$.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Scheme of the proof of Theorem 2 (continued)

Let x^* — be the center of the «cube» B, i.e., $x^* = \hat{x} + \frac{1}{2p} e$. Consider the function $\bar{f}(x) = \min\{0, M\|x - x^*\|_\infty - \varepsilon\}$. The function $\bar{f}(x)$ is Lipschitz with constant M with respect to the ℓ_∞-norm and takes its minimum value $-\varepsilon$ at the point x^*.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Scheme of the proof of Theorem 2 (continued)

Let x^* — be the center of the «cube» B, i.e., $x^* = \hat{x} + \frac{1}{2p}e$. Consider the function $\bar{f}(x) = \min\{0, M\|x - x^*\|_\infty - \varepsilon\}$. The function $\bar{f}(x)$ is Lipschitz with constant M with respect to the ℓ_∞-norm and takes its minimum value $-\varepsilon$ at the point x^*. Moreover, the function $\bar{f}(x)$ is different from zero only inside the cube $B' = \{x \mid \|x - x^*\| \leq \frac{\varepsilon}{M}\}$, which lies inside the cube B, since $2p \leq \frac{M}{\varepsilon}$.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Scheme of the proof of Theorem 2 (continued)

Let x^* — be the center of the «cube» B, i.e., $x^* = \hat{x} + \frac{1}{2p}e$. Consider the function $\bar{f}(x) = \min\{0, M\|x - x^*\|_\infty - \varepsilon\}$. The function $\bar{f}(x)$ is Lipschitz with constant M with respect to the ℓ_∞-norm and takes its minimum value $-\varepsilon$ at the point x^*. Moreover, the function $\bar{f}(x)$ is different from zero only inside the cube $B' = \{x \mid \|x - x^*\| \leq \frac{\varepsilon}{M}\}$, which lies inside the cube B, since $2p \leq \frac{M}{\varepsilon}$. Hence, the considered method cannot find a ε-solution on this function. Contradiction.

Complexity of optimization problems. Class of problems of minimization of Lipschitz functions

Scheme of the proof of Theorem 2 (continued)

Let x^* — be the center of the «cube» B, i.e., $x^* = \hat{x} + \frac{1}{2p}e$. Consider the function $\bar{f}(x) = \min\{0, M\|x - x^*\|_\infty - \varepsilon\}$. The function $\bar{f}(x)$ is Lipschitz with constant M with respect to the ℓ_∞-norm and takes its minimum value $-\varepsilon$ at the point x^*. Moreover, the function $\bar{f}(x)$ is different from zero only inside the cube $B' = \{x \mid \|x - x^*\| \leq \frac{\varepsilon}{M}\}$, which lies inside the cube B, since $2p \leq \frac{M}{\varepsilon}$. Hence, the considered method cannot find a ε-solution on this function. Contradiction.

Thus, in the our class of problems any method has rather pessimistic estimates for the convergence rate. The question arises: what properties should be required from the class of optimized functions to make the estimates more optimistic?

Convex and smooth functions

Good news:

- A rich and interesting theory
- There are efficient algorithms for finding approximate solutions.

Convex and smooth functions

Good news:

- A rich and interesting theory
- There are efficient algorithms for finding approximate solutions.

Bad news:

- The class of convex and smooth problems is not very broad
- In practice we often have to face non-convex problems

Convex and smooth functions

- Nevertheless, sometimes convex smooth optimization methods also perform well in practice on nonconvex or nonsmooth problems that are locally convex or smooth.

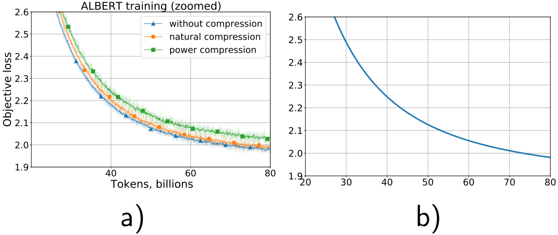


Figure: a) ALBERT (big NLP) model training: loss function, b) rate $1/k^2$ – theoretical rate of GD with momentum for smooth and convex problems

- The theory for convex and smooth problems allows us to compare methods and determine which of them is better at all/better suited for a certain class of problems.

Convex sets

Definition 1

A set $Q \subseteq \mathbb{R}^d$ is called convex if for any two points $x, y \in Q$ and for any number $\alpha \in [0, 1]$ the point $z = \alpha x + (1 - \alpha) y$ belongs to the set Q.

Convex sets

Definition 1

A set $Q \subseteq \mathbb{R}^d$ is called convex if for any two points $x, y \in Q$ and for any number $\alpha \in [0, 1]$ the point $z = \alpha x + (1 - \alpha) y$ belongs to the set Q.

This means that along with any two points, the set contains a segment connecting them.

Convex sets. Examples

- $Q = \mathbb{R}^d$

Convex sets. Examples

- $Q = \mathbb{R}^d$ — convex set.
- $Q = \{x \in \mathbb{R}^d \mid x_i \geq 0,\ i = 1, \ldots, d\}$

Convex sets. Examples

- $Q = \mathbb{R}^d$ — convex set.
- $Q = \{x \in \mathbb{R}^d \mid x_i \geq 0,\ i = 1, \ldots, d\}$ — convex set. Indeed, let $x, y \in Q$, $\alpha \in [0, 1]$ and $z = \alpha x + (1 - \alpha)y$. Since $\alpha \geq 0$ and $1 - \alpha \geq 0$, we get: $z_i = \alpha x_i + (1 - \alpha)y_i \geq 0$ for all $i = 1, \ldots, d$, and then, $z \in Q$.
- $Q = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$

Convex sets. Examples

- $Q = \mathbb{R}^d$ — convex set.
- $Q = \{x \in \mathbb{R}^d \mid x_i \geq 0,\ i = 1, \dots, d\}$ — convex set. Indeed, let $x, y \in Q$, $\alpha \in [0, 1]$ and $z = \alpha x + (1 - \alpha) y$. Since $\alpha \geq 0$ and $1 - \alpha \geq 0$, we get: $z_i = \alpha x_i + (1 - \alpha) y_i \geq 0$ for all $i = 1, \dots, d$, and then, $z \in Q$.
- $Q = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$ — convex set. We have: $\|z\| = \|\alpha x + (1 - \alpha) y\| \leq \|\alpha x\| + \|(1 - \alpha) y\| = \alpha \|x\| + (1 - \alpha) \|y\|$

Convex sets. Examples

- $Q = \mathbb{R}^d$ — convex set.
- $Q = \{x \in \mathbb{R}^d \mid x_i \geq 0, \; i = 1, \dots, d\}$ — convex set. Indeed, let $x, y \in Q$, $\alpha \in [0, 1]$ and $z = \alpha x + (1 - \alpha)y$. Since $\alpha \geq 0$ and $1 - \alpha \geq 0$, we get: $z_i = \alpha x_i + (1 - \alpha)y_i \geq 0$ for all $i = 1, \dots, d$, and then, $z \in Q$.
- $Q = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$ — convex set. We have: $\|z\| = \|\alpha x + (1 - \alpha)y\| \leq \|\alpha x\| + \|(1 - \alpha)y\| = \alpha\|x\| + (1 - \alpha)\|y\| \leq \alpha + (1 - \alpha) = 1.$
- $Q = \{x \in \mathbb{R}^d \mid \|x\| = 1\}$

Convex sets. Examples

- $Q = \mathbb{R}^d$ — convex set.
- $Q = \{x \in \mathbb{R}^d \mid x_i \geq 0, \ i = 1, \dots, d\}$ — convex set. Indeed, let $x, y \in Q$, $\alpha \in [0,1]$ and $z = \alpha x + (1-\alpha)y$. Since $\alpha \geq 0$ and $1 - \alpha \geq 0$, we get: $z_i = \alpha x_i + (1-\alpha)y_i \geq 0$ for all $i = 1, \dots, d$, and then, $z \in Q$.
- $Q = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$ — convex set. We have: $\|z\| = \|\alpha x + (1-\alpha)y\| \leq \|\alpha x\| + \|(1-\alpha)y\| = \alpha\|x\| + (1-\alpha)\|y\| \leq \alpha + (1-\alpha) = 1$.
- $Q = \{x \in \mathbb{R}^d \mid \|x\| = 1\}$ — non-convex set. Let us show: take an arbitrary vector $x \in Q$. Тогда $-x \in Q$, since $\|x\| = \|-x\| = 1$. However, any nontrivial convex combination of x and $-x$ does not lie in Q: if $\alpha \in (0,1)$, then $\|\alpha x + (1-\alpha)(-x)\| = \|(2\alpha - 1)x\| = |2\alpha - 1| < 1$.

Convex functions

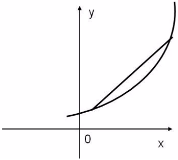
Definition 2

A function $f(x)$ defined on a **convex** set $Q \subseteq \mathbb{R}^d$ is called **convex** if for any two points $x, y \in Q$ and for any number $\alpha \in [0, 1]$ the following inequality holds:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y). \tag{11}$$

(if we have sign $<$ instead of $\leq$ for all $x \neq y$, $\alpha \in (0, 1)$, then the function is **strictly convex**)

In the one-dimensional case, this means that between any two points $x, y \in Q$, the graph of the function f passes not above the segment connecting $f(x)$ and $f(y)$.



Convex functions. Examples

- $f(x) = \langle a, x \rangle$

Convex functions. Examples

- $f(x) = \langle a, x \rangle$ — convex function. Since $f(\alpha x + (1-\alpha)y) = \alpha\langle a, x\rangle + (1-\alpha)\langle a, y\rangle = \alpha f(x) + (1-\alpha) f(y)$ for all $\alpha \in [0,1]$.
- $f(x) = \|x\|_2^2$

Convex functions. Examples

- $f(x) = \langle a, x \rangle$ — convex function. Since $f(\alpha x + (1-\alpha)y) = \alpha\langle a, x\rangle + (1-\alpha)\langle a, y\rangle = \alpha f(x) + (1-\alpha)f(y)$ for all $\alpha \in [0,1]$.
- $f(x) = \|x\|_2^2$ — convex function. Since for all $\alpha \in [0,1]$

$$f(\alpha x + (1-\alpha)y) \quad = \quad \alpha^2\|x\|_2^2 + 2\alpha(1-\alpha)\langle x, y\rangle + (1-\alpha)^2\|y\|_2^2$$

Convex functions. Examples

- $f(x) = \langle a, x \rangle$ — convex function. Since $f(\alpha x + (1-\alpha)y) = \alpha\langle a, x\rangle + (1-\alpha)\langle a, y\rangle = \alpha f(x) + (1-\alpha) f(y)$ for all $\alpha \in [0,1]$.
- $f(x) = \|x\|_2^2$ — convex function. Since for all $\alpha \in [0,1]$

$$\begin{aligned} f(\alpha x + (1-\alpha)y) &= \alpha^2\|x\|_2^2 + 2\alpha(1-\alpha)\langle x, y\rangle + (1-\alpha)^2\|y\|_2^2 \\ &\overset{\text{K.-B.}}{\leq} (\alpha^2 + \alpha(1-\alpha))\|x\|_2^2 \\ &\quad +((1-\alpha)^2 + \alpha(1-\alpha))\|y\|_2^2 \end{aligned}$$

Convex functions. Examples

- $f(x) = \langle a, x \rangle$ — convex function. Since $f(\alpha x + (1-\alpha)y) = \alpha\langle a, x\rangle + (1-\alpha)\langle a, y\rangle = \alpha f(x) + (1-\alpha) f(y)$ for all $\alpha \in [0,1]$.
- $f(x) = \|x\|_2^2$ — convex function. Since for all $\alpha \in [0,1]$

$$
\begin{aligned}
f(\alpha x + (1-\alpha)y) &= \alpha^2\|x\|_2^2 + 2\alpha(1-\alpha)\langle x, y\rangle + (1-\alpha)^2\|y\|_2^2 \\
&\overset{\text{K.-B.}}{\leq} (\alpha^2 + \alpha(1-\alpha))\|x\|_2^2 \\
&\qquad + ((1-\alpha)^2 + \alpha(1-\alpha))\|y\|_2^2 \\
&= \alpha f(x) + (1-\alpha) f(y),
\end{aligned}
$$

  where K.-B. means that the transition is fair by virtue of the Cauchy-Bunyakovsky-Shwartz inequality.
- $f(x) = \langle a, x \rangle - \|x\|^2$

Convex functions. Examples

- $f(x) = \langle a, x\rangle$ — convex function. Since $f(\alpha x + (1-\alpha)y) = \alpha\langle a, x\rangle + (1-\alpha)\langle a, y\rangle = \alpha f(x) + (1-\alpha)f(y)$ for all $\alpha \in [0, 1]$.
- $f(x) = \|x\|_2^2$ — convex function. Since for all $\alpha \in [0, 1]$

$$
\begin{aligned}
f(\alpha x + (1-\alpha)y) &= \alpha^2\|x\|_2^2 + 2\alpha(1-\alpha)\langle x, y\rangle + (1-\alpha)^2\|y\|_2^2 \\
&\overset{\text{K.-B.}}{\leq} (\alpha^2 + \alpha(1-\alpha))\|x\|_2^2 \\
&\quad\quad +((1-\alpha)^2 + \alpha(1-\alpha))\|y\|_2^2 \\
&= \alpha f(x) + (1-\alpha)f(y),
\end{aligned}
$$

  where K.-B. means that the transition is fair by virtue of the Cauchy-Bunyakovsky-Shwartz inequality.
- $f(x) = \langle a, x\rangle - \|x\|^2$ — non-convex function.

Strongly convex functions

Definition 3

A function $f(x)$ defined on a convex set $Q \subseteq \mathbb{R}^d$ is called μ strongly convex if for any two points $x, y \in Q$ and for any number $\alpha \in [0,1]$ the following inequality holds:

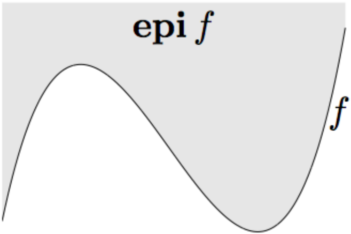
$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \alpha(1-\alpha)\frac{\mu}{2}\|x-y\|_2^2. \quad (12)$$

Properties of convex functions I

1. Convex functions are continuous at all interior points of the domain of definition.
2. A strongly convex function is obviously strictly convex. The converse is not true.
3. A function is convex if and only if its epigraph is a convex set, where by the epigraph of a function f defined on the set $Q \subseteq \mathbb{R}^d$ we mean the following set:

$$\mathrm{epi} f = \{(x, t) \mid x \in Q,\ t \in \mathbb{R},\ t \geq f(x)\} \subseteq \mathbb{R}^{d+1}$$

Properties of convex functions II



4. If $f(x)$ is a convex function, then the set of

$$\mathrm{C}_\gamma = \{x \in Q \subseteq \mathbb{R}^d \mid f(x) \leq \gamma,\ \gamma = \text{const}\}$$

is convex.

Properties of convex functions III

5. For a convex function $f(x)$ the following is true **Jensen's inequality**

$$f\left(\sum_{i=1}^{n}\lambda_i x_i\right) \leq \sum_{i=1}^{n}\lambda_i f(x_i)$$

for all $\lambda_i \geq 0$ и $\sum\limits_{i=1}^{n}\lambda_i = 1$.
Jensen's inequality generalizes to the case of a convex combination of an infinite (countable or uncountable) number of points.

6. If $p(x) \geq 0,\ \int_{\mathrm{S}} p(x)dx = 1,\ \mathrm{S} \subseteq \mathrm{dom} f$

$$f\left(\int\limits_{\mathrm{S}} p(x)xdx\right) \leq \int\limits_{\mathrm{S}} f(x)p(x)dx.$$

Properties of convex functions IV

7. If x is a random variable and $x \in \mathrm{dom} f$ with probability 1, then

$$f(\mathbb{E}x) \leq \mathbb{E}[f(x)],$$

assuming the mathematical expectation exists.

Operations that preserve convexity I

1. Positive weighted sum

Let $f_1, ..., f_m$ be convex functions on the convex set G; $\lambda_1, ..., \lambda_m$ are non-negative numbers. Then the function

$$f(x) = \sum_{i=1}^{m} \lambda_i f_i(x) \quad \text{– convex on } \mathrm{G}.$$

Proof. Let us prove by definition. Let $x, y \in \mathrm{G}$, $\alpha \in [0, 1]$:

$$f(\alpha x + (1-\alpha)y) = \sum_{i=1}^{m} \lambda_i f_i(\alpha x + (1-\alpha)y) \leq \sum_{i=1}^{m} \lambda_i \left[\alpha f_i(x) + (1-\alpha) f_i(y)\right]$$

$$= \alpha \sum_{i=1}^{m} \lambda_i f_i(x) + (1-\alpha) \sum_{i=1}^{m} \lambda_i f_i(y) = \alpha f(x) + (1-\alpha) f(y)$$

Operations that preserve convexity II

2. Maximum of convex functions

Let $f_1, ..., f_m$ be convex functions on a convex set G. Then the function

$$f(x) = \max_{i=\overline{1,m}} f_i(x) \quad \text{– convex on } \mathrm{G}.$$

Proof. Let us prove by definition. Let $x, y \in \mathrm{G}$, $\alpha \in [0,1]$:

$$f(\alpha x + (1-\alpha)y) = \max\{\ f_1(\alpha x + (1-\alpha)y),\ f_2(\alpha x + (1-\alpha)y)\ \} \leq$$

$$\max\{\ \alpha f_1(x) + (1-\alpha) f_1(y),\ \alpha f_2(x) + (1-\alpha) f_2(y)\ \} \leq$$

$$\alpha \max\{\ f_1(x), f_2(x)\ \} + (1-\alpha)\max\{\ f_1(y), f_2(y)\ \} = \alpha f(x) + (1-\alpha) f(y)$$

Operations that preserve convexity III

3. Pointwise supremum.

If a function of two arguments $g(x, y)$ is convex on $x \in \mathbb{R}^n$ for any $y \in \mathrm{Y} \subseteq \mathbb{R}^m$, then the function

$$f(x) = \sup_{y \in \mathrm{Y}} g(x, y) \quad \text{– convex on } x.$$

Operations that preserve convexity IV

Example

Maximum eigenvalue of a symmetric matrix

$$f(X) = \lambda_{\max}(X), \ \mathrm{dom} f = \mathbb{S}^m \quad \text{– convex.}$$

Proof.

$f(X)$ can be represented as

$$f(X) = \sup \left\{ y^\top X y \mid \|y\|_2 = 1 \right\},$$

i.e. as a pointwise supremum of the family of linear functions from X. □

Operations that preserve convexity V

4. Affine argument substitution

- Let $\varphi(x)$ be a convex function on a convex set $\mathrm{G} \subseteq \mathbb{R}^m$.
- Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $\mathrm{X} = \{x \in \mathbb{R}^n \mid Ax + b \in \mathrm{G}\}$, $\mathrm{X} \neq \varnothing$.
- Then $f(x) = \varphi(Ax + b)$ - convex function on X.

Proof.

We prove by definition, let $x, y \in \mathrm{X}$, $\alpha \in [0, 1]$.

$$f(\alpha x + (1-\alpha)y) = \varphi\left(A(\alpha x + (1-\alpha)y) + b\right)$$

$$\varphi\left(\alpha(Ax + b) + (1-\alpha)(Ay + b)\right) \leq \alpha\varphi(Ax + b) + (1-\alpha)\varphi(Ay + b)$$
$$= \alpha f(x) + (1-\alpha) f(y)$$

$\square$

Operations that preserve convexity VI

5. Superposition of convex functions

- Let $h(x)$ be a convex function, $P(y)$ be convex and non-decreasing. Then $f(x) = P(h(x))$ is a convex function.

Proof.

We prove by definition, let $x, y \in \mathrm{X}$, $\alpha \in [0,1]$.

$$f(\alpha x + (1-\alpha)y) = P(h(\alpha x + (1-\alpha)y)) \leq P(\alpha h(x) + (1-\alpha)h(y))$$

$$\leq \alpha P(h(x)) + (1-\alpha)P(h(y)) \leq \alpha f(x) + (1-\alpha)f(y)$$

$\square$

Operations that preserve convexity VII

Proof

Is the function

$$f(x) = \exp\left(\sum_{i=1}^{m} |a_i^\top x - b_i|\right)$$

convex?

- $g_i(x) = |a_i^\top x - b_i|$ - is convex, since the absolute value $|\cdot|$ is a convex function and the function $a_i^\top x - b_i$ is linear, that is, an affine substitution of the argument
- $h(x) = \sum_{i=1}^{m} g_i(x)$ - is convex as the sum of convex functions
- $f(x) = P(h(x))$ - is convex as a superposition of convex, non-decreasing and convex functions

Smooth optimization problems

Definition 4

A differentiable function $f(x)$ defined on the set $Q \subseteq \mathbb{R}^d$ is called L-smooth if the following inequality holds for any two points $x, y \in Q$:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2. \tag{13}$$

Smooth optimization problems

Definition 4

A differentiable function $f(x)$ defined on the set $Q \subseteq \mathbb{R}^d$ is called L-smooth if the following inequality holds for any two points $x, y \in Q$:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2. \tag{13}$$

Examples

- $f(x) = \frac{1}{2}\|x\|_2^2$

Smooth optimization problems

Definition 4

A differentiable function $f(x)$ defined on the set $Q \subseteq \mathbb{R}^d$ is called L-smooth if the following inequality holds for any two points $x, y \in Q$:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2. \tag{13}$$

Examples

- $f(x) = \frac{1}{2}\|x\|_2^2$ — 1-smooth function on $\mathbb{R}^d$.
- $f(x) = \|x\|_2^3$

Smooth optimization problems

Definition 4

A differentiable function $f(x)$ defined on the set $Q \subseteq \mathbb{R}^d$ is called L-smooth if the following inequality holds for any two points $x, y \in Q$:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2. \tag{13}$$

Examples

- $f(x) = \frac{1}{2}\|x\|_2^2$ — 1-smooth function on $\mathbb{R}^d$.
- $f(x) = \|x\|_2^3$ is not L-smooth on $\mathbb{R}^d$ for any L.
- $f(x) = x^3$

Smooth optimization problems

Definition 4

A differentiable function $f(x)$ defined on the set $Q \subseteq \mathbb{R}^d$ is called L-smooth if the following inequality holds for any two points $x, y \in Q$:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2. \tag{13}$$

Examples

- $f(x) = \frac{1}{2}\|x\|_2^2$ — 1-smooth function on $\mathbb{R}^d$.
- $f(x) = \|x\|_2^3$ is not L-smooth on $\mathbb{R}^d$ for any L.
- $f(x) = x^3$ — 12-is a smooth function on the segment $[1, 2]$, which follows from Lagrange's mean and boundedness theorem $f''(x) = 6x$ on the segment $[1, 2]$.