Optimality conditions. KKT

Background

Extreme value (Weierstrass) theorem

Let $S \subset \mathbb{R}^n$ be compact set and $f(x)$ continuous function on S. So that, the point of the global minimum of the function $f(x)$ on S exists.

Lagrange multipliers

Consider simple yet practical case of equality constraints:

$$
\begin{split}
& f(x) \to \min\limits_{x \in \mathbb{R}^n} \\
\text{s.t. } & h_i(x) = 0, i = 1, \ldots, m
\end{split}
$$

The basic idea of Lagrange method implies switch from conditional to unconditional optimization through increasing the dimensionality of the problem:

$$
L(x, \lambda) = f(x) + \sum\limits_{i=1}^m \lambda_i h_i(x) \to \min\limits_{x \in \mathbb{R}^n, \lambda \in \mathbb{R}^m}
$$

General formulations and conditions

$$
f(x) \to \min\limits_{x \in S}
$$

We say that the problem has a solution if the following set **is not empty**: $x^* \in S$, in which the minimum or the infimum of the given function is achieved.

Unconstrained optimization

General case

Let $f(x) : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function.

$$
f(x) \to \min\limits_{x \in \mathbb{R}^n} \tag{UP}
$$

If x^* - is a local minimum of $f(x)$, then:

$$\nabla f(x^*) = 0 \tag{UP:Necessary}$$

If $f(x)$ at some point x^* satisfies the following conditions:

$$H_f(x^*) = \nabla^2 f(x^*) \succeq (\preceq) 0, \tag{UP:Sufficient}$$

then (if necessary condition is also satisfied) x^* is a local minimum(maximum) of $f(x)$.

Convex case

It should be mentioned, that in **convex** case (i.e., $f(x)$ is convex) necessary condition becomes sufficient. Moreover, we can generalize this result on the class of non-differentiable convex functions.

Let $f(x) : \mathbb{R}^n \to \mathbb{R}$ - convex function, then the point x^* is the solution of (UP) if and only if:

$$0_n \in \partial f(x^*)$$

One more important result for convex constrained case sounds as follows. If $f(x) : S \to \mathbb{R}$ - convex function defined on the convex set S, then:

- Any local minima is the global one.
- The set of the local minimizers S^* is convex.
- If $f(x)$ - strongly convex function, then S^* contains only one single point $S^* = x^*$.

Optimization with equality conditions

Intuition

Things are pretty simple and intuitive in unconstrained problem. In this section we will add one equality constraint, i.e.

$$\begin{split} & f(x) \to \min_{x \in \mathbb{R}^n} \\ \text{s.t. } & h(x) = 0 \end{split}$$

We will try to illustrate approach to solve this problem through the simple example with $f(x) = x_1 + x_2$ and $h(x) = x_1^2 + x_2^2 - 2$

Generally: in order to move from x_F along the budget set towards decreasing the function, we need to guarantee two conditions:

$$\langle \delta x, \nabla h(x_F) \rangle = 0$$

$$\langle \delta x, -\nabla f(x_F) \rangle > 0$$

Let's assume, that in the process of such a movement we have come to the point where

$$\nabla f(x) = \lambda \nabla h(x)$$

$$\langle \delta x, -\nabla f(x) \rangle = -\langle \delta x, \lambda \nabla h(x) \rangle = 0$$

Then we came to the point of the budget set, moving from which it will not be possible to reduce our function. This is the local minimum in the limited problem :)

So let's define a Lagrange function (just for our convenience):

$$L(x, \lambda) = f(x) + \lambda h(x)$$

Then the point x^* be the local minimum of the problem described above, if and only if:

$$\begin{split} & \nabla_x L(x^*, \lambda^*) = 0 \text{ that's written above}\\ & \nabla_\lambda L(x^*, \lambda^*) = 0 \text{ condition of being in budget set} \\ & \langle y, \nabla^2_{xx} L(x^*, \lambda^*) y \rangle \geq 0, \quad \forall y \in \mathbb{R}^n : \nabla h(x^*)^\top y = 0 \end{split}$$

We should notice that $L(x^*, \lambda^*) = f(x^*)$.

General formulation

$$\begin{split} & f(x) \to \min\limits_{x \in \mathbb{R}^n} \\ \text{s.t. } & h_i(x) = 0, \; i = 1, \ldots, m \end{split}$$

Solution

$$L(x, \lambda) = f(x) + \sum\limits_{i=1}^m \lambda_i h_i(x) = f(x) + \lambda^\top h(x)$$

Let $f(x)$ and $h_i(x)$ be twice differentiable at the point x^* and continuously differentiable in some neighborhood x^*. The local minimum conditions for $x \in \mathbb{R}^n, \lambda \in \mathbb{R}^m$ are written as

$$\begin{split} & \nabla_x L(x^*, \lambda^*) = 0 \\ & \nabla_\lambda L(x^*, \lambda^*) = 0 \\ & \langle y, \nabla^2_{xx} L(x^*, \lambda^*) y \rangle \geq 0, \quad \forall y \in \mathbb{R}^n : \nabla h(x^*)^\top y = 0 \end{split}$$

Depending on the behavior of the Hessian, the critical points can have a different character.

Optimization with inequality conditions

Example

$$f(x) = x_1^2 + x_2^2 \quad g(x) = x_1^2 + x_2^2 - 1$$

$$\begin{split} & f(x) \to \min\limits_{x \in \mathbb{R}^n} \\ \text{s.t. } & g(x) \leq 0 \end{split}$$

Thus, if the constraints of the type of inequalities are inactive in the UM problem, then don't worry and write out the solution to the UM problem. However, this is not a heal-all :) Consider the second childish example

$$f(x) = (x_1 - 1.1)^2 + (x_2 + 1.1)^2 \quad g(x) = x_1^2 + x_2^2 - 1$$

$$\begin{split} & f(x) \to \min_{x \in \mathbb{R}^n} \\ \text{s.t. } & g(x) \leq 0 \end{split}$$

So, we have a problem:

$$\begin{split} & f(x) \to \min_{x \in \mathbb{R}^n} \\ \text{s.t. } & g(x) \leq 0 \end{split}$$

Two possible cases:

1. $$\begin{split} & g(x^*) < 0 \\ & \nabla f(x^*) = 0 \\ & \nabla^2 f(x^*) > 0 \end{split}$$
2. $$\begin{split} & g(x^*) = 0 \\ & - \nabla f(x^*) = \mu \nabla g(x^*), \quad \mu > 0 \\ & \langle y, \nabla^2_{xx} L(x^*, \mu^*) y \rangle \geq 0, \quad \forall y \in \mathbb{R}^n : \nabla g(x^*)^\top y = 0 \end{split}$$

Combining two possible cases, we can write down the general conditions for the problem:

$$\begin{split} & f(x) \to \min_{x \in \mathbb{R}^n} \\ \text{s.t. } & g(x) \leq 0 \end{split}$$

Let's define the Lagrange function:

$$L(x, \mu) = f(x) + \mu g(x)$$

Then x^* point - local minimum of the problem described above, if and only if:

$$\begin{split} & (1) \ \nabla_x L(x^*, \mu^*) = 0 \\ & (2) \ \mu^* \geq 0 \\ & (3) \ \mu^* g(x^*) = 0 \\ & (4) \ g(x^*) \leq 0 \\ & (5) \ \langle y, \nabla^2_{xx} L(x^*, \mu^*) y \rangle \geq 0, \quad \forall y \in \mathbb{R}^n : \nabla g(x^*)^\top y = 0 \end{split}$$

It's noticeable, that $L(x^*, \mu^*) = f(x^*)$. Conditions $\mu^* = 0, (1), (4)$ are the first scenario realization, and conditions $\mu^* > 0, (1), (3)$ - the second.

General formulation

$$
\begin{split}
& f(x) \to \min_{x \in \mathbb{R}^n} \\
\text{s.t. } & g_i(x) \leq 0, \; i = 1, \ldots, m \\
& h_j(x) = 0, \; j = 1, \ldots, p
\end{split}
$$

This formulation is a general problem of mathematical programming. From now, we only consider $\textbf{regular}$ tasks. This is a very important remark from a formal point of view. Those wishing to understand in more detail, please refer to Google.

Solution

$$
L(x, \mu, \lambda) = f(x) + \sum\limits_{j=1}^p \lambda_j h_j(x) + \sum\limits_{i=1}^m \mu_i g_i(x)
$$

Karush-Kuhn-Tucker conditions

Let x^* be a solution to a mathematical programming problem, and the functions f, h_j, g_i are differentiable. Then there are λ^* and μ^* such that the following conditions are carried out:

- $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$
- $\nabla_\lambda L(x^*, \lambda^*, \mu^*) = 0$
- $\mu^*_j \geq 0$
- $\mu^*_j g_j(x^*) = 0$
- $g_j(x^*) \leq 0$

These conditions are sufficient if the problem is regular, i.e. if: 1) the given problem is a convex optimization problem (i.e., the functions f and g_i are convex, h_i are affine) and the Slater condition is satisfied; or 2) strong duality is fulfilled.

References

- [Lecture](#) on KKT conditions (very intuitive explanation) in course "Elements of Statistical Learning" @ KTH.
- [One-line proof of KKT](#)

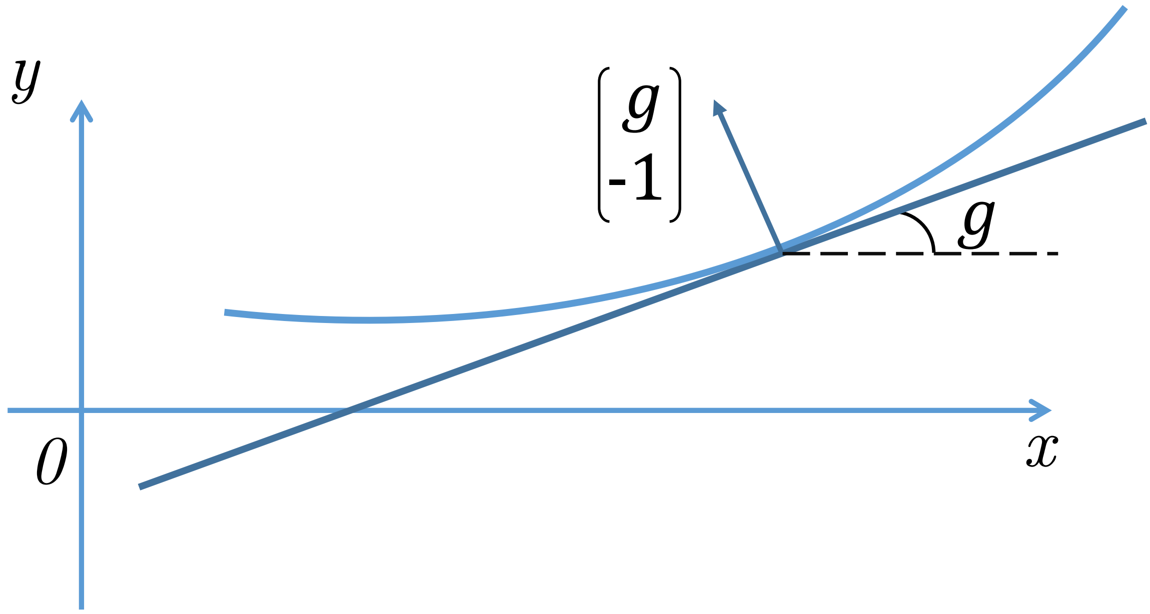
Subgradient and subdifferential

Motivation

Важным свойством непрерывной выпуклой функции $f(x)$ является то, что в выбранной точке x_0 для всех $x \in \text{dom } f$ выполнено неравенство:

$$
f(x) \geq f(x_0) + \langle g, x - x_0 \rangle
$$

для некоторого вектора g, то есть касательная к графику функции является *глобальной* оценкой снизу для функции.



- Если $f(x)$ - дифференцируема, то $g = \nabla f(x_0)$
- Не все непрерывные выпуклые функции дифференцируемы :)

Не хочется лишаться такого вкусного свойства.

Subgradient

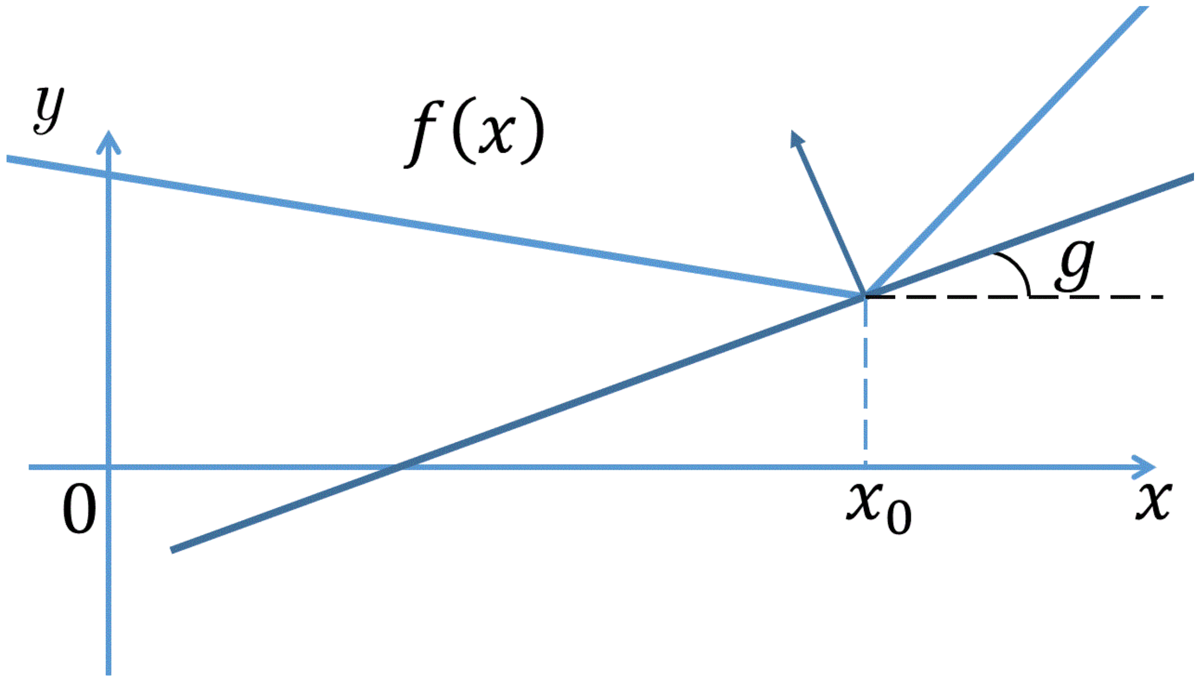
Вектор g называется **субградиентом** функции $f(x) : S \to \mathbb{R}$ в точке x_0, если $\forall x \in S$:

$$
f(x) \geq f(x_0) + \langle g, x - x_0 \rangle
$$

Subdifferential

Множество всех субградиентов функции $f(x)$ в точке x_0 называется **субдифференциалом** f в x_0 и обозначается $\partial f(x_0)$.

- Если $x_0 \in \mathbf{ri} S$, то $\partial f(x_0)$ выпуклое компактное множество.
- Выпуклая функция $f(x)$ дифференцируема в точке $x_0 \iff \partial f(x_0) = \nabla f(x_0)$
- Если $\partial f(x_0) \neq \emptyset \quad \forall x_0 \in S$, то $f(x)$ - выпукла на S.



Moreau - Rockafellar theorem (subdifferential of a linear combination)

Пусть $f_i(x)$ - выпуклые функции на выпуклых множествах $S_i, \; i = \overline{1,n}$.
Тогда, если $\bigcap\limits_{i=1}^n \mathbf{ri} S_i \neq \emptyset$ то функция $f(x) = \sum\limits_{i=1}^n a_i f_i(x), \; a_i > 0$ имеет субдифференциал $\partial_S f(x)$ на множестве $S = \bigcap\limits_{i=1}^n S_i$ и

$$
\partial_S f(x) = \sum_{i=1}^n a_i \partial_{S_i} f_i(x)
$$

Dubovitsky - Milutin theorem (subdifferential of a point-wise maximum)

Пусть $f_i(x)$ - выпуклые функции на открытом выпуклом множестве $S \subseteq \mathbb{R}^n, \; x_0 \in S$, а поточечный максимум определяется как $f(x) = \max\limits_i f_i(x)$. Тогда:

$$
\partial_S f(x_0) = \mathbf{conv} \left\{ \bigcup_{i \in I(x_0)} \partial_S f_i(x_0) \right\},
$$

где $I(x) = \{ i \in [1:m] : f_i(x) = f(x) \}$

Chain rule for subdifferentials

Пусть $g_1, \ldots, g_m$ - выпуклые функции на открытом выпуклом множестве $S \subseteq \mathbb{R}^n$, $g = (g_1, \ldots, g_m)$ - образованная из них вектор - функция, φ - монотонно неубывающая выпуклая функция на открытом выпуклом множестве $U \subseteq \mathbb{R}^m$, причем $g(S) \subseteq U$. Тогда субдифференциал функции $f(x) = \varphi\left(g(x)\right)$ имеет вид:

$$
\partial f(x) = \bigcup_{p \in \partial \varphi(u)} \left(\sum_{i=1}^m p_i \partial g_i(x) \right),
$$

где $u = g(x)$

В частности, если функция φ дифференцируема в точке $u = g(x)$, то формула запишется так:

$$
\partial f(x) = \sum_{i=1}^m \dfrac{\partial \varphi}{\partial u_i}(u) \partial g_i(x)
$$

Subdifferential calculus

- $\partial (\alpha f)(x) = \alpha \partial f(x)$, for $\alpha \geq 0$
- $\partial (\sum f_i)(x) = \sum \partial f_i (x)$, f_i - выпуклые функции
- $\partial (f(Ax + b))(x) = A^T\partial f(Ax + b)$, f - выпуклая функция

Examples

Концептуально, различают три способа решения задач на поиск субградиента:

- Теоремы Моро - Рокафеллара, композиции, максимума
- Геометрически
- По определению

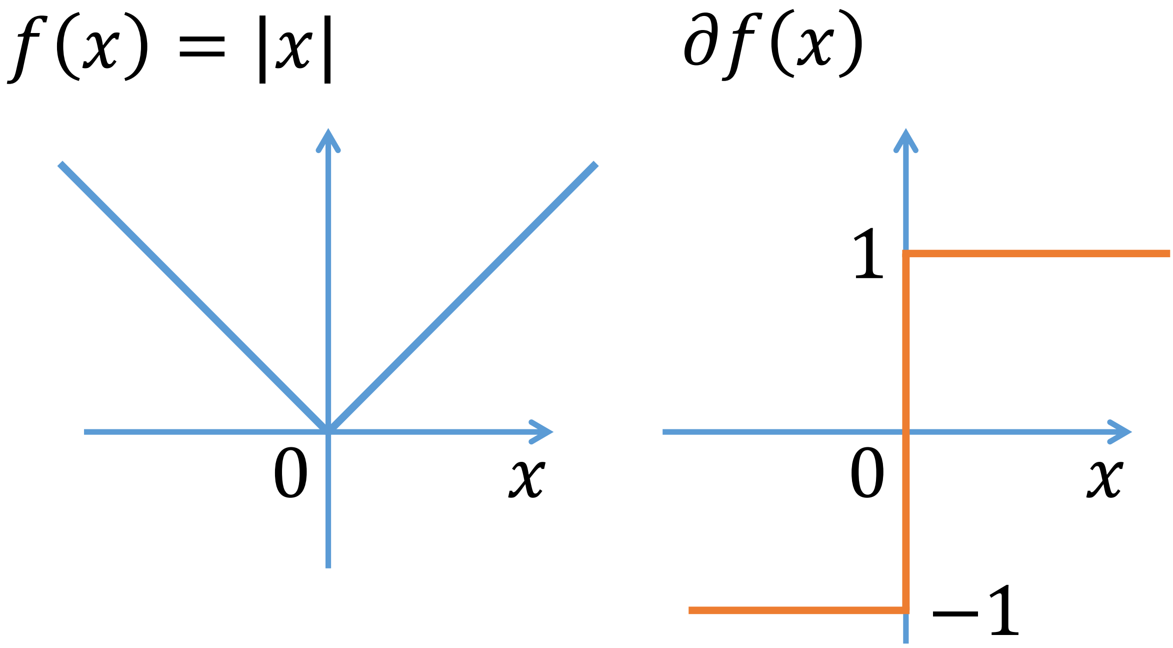
1

Найти $\partial f(x)$, если $f(x) = |x|$

Решение:

Решить задачу можно либо геометрически (в каждой точке числовой прямой указать угловые коэффициенты прямых, глобально подпирающих функцию снизу), либо по теореме Моро - Рокафеллара, рассмотрев $f(x)$ как композицию выпуклых функций:

$$f(x) = \max\{-x, x\}$$



2

Найти $\partial f(x)$, если $f(x) = |x-1| + |x+1|$

Решение:

Совершенно аналогично применяем теорему Моро - Рокафеллара, учитывая следующее:

$$\partial f_1(x) = \begin{cases} -1, & x < 1 \\ [-1;1], & x = 1 \\ 1, & x > 1 \end{cases} \qquad \partial f_2(x) = \begin{cases} -1, & x < -1 \\ [-1;1], & x = -1 \\ 1, & x > -1 \end{cases}$$

Таким образом:

$$\partial f(x) = \begin{cases} -2, & x < -1 \\ [-2;0], & x = -1 \\ 0, & -1 < x < 1 \\ [0;2], & x = 1 \\ 2, & x > 1 \end{cases}$$

3

Найти $\partial f(x)$, если $f(x) = [\max(0, f_0(x))]^q$. Здесь $f_0(x)$ - выпуклая функция на открытом выпуклом множестве S, $q \geq 1$.

Решение:

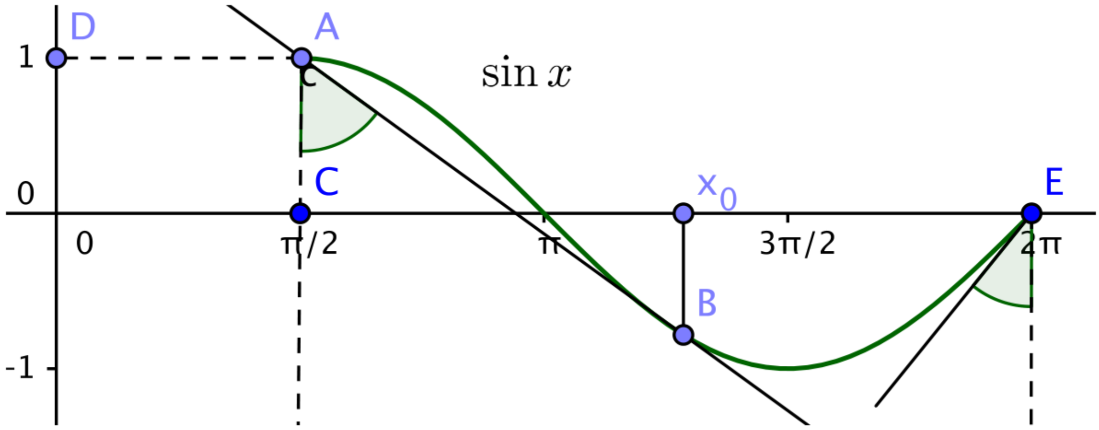
Согласно теореме о композиции (функция $\varphi(x) = x^q$ - дифференцируема), а $g(x) = \max(0, f_0(x))$ имеем: $\partial f(x) = q(g(x))^{q-1} \partial g(x)$

По теореме о поточечном максимуме:

$$\partial g(x) = \begin{cases} \partial f_0(x), & f_0(x) > 0, \\ \{0\}, & f_0(x) < 0 \\ \{a \mid a = \lambda a',\ 0 \le \lambda \le 1,\ a' \in \partial f_0(x)\}, & f_0(x) = 0 \end{cases}$$

4

Найти $\partial f(x)$, если $f(x) = \sin x, x \in [\pi/2; 2\pi]$



$$\partial f_G(x) = \begin{cases} (-\infty, \cos x_0], & x = \pi/2; \\ \varnothing, & x \in (\pi/2, x_0); \\ \cos x, & x \in [x_0, 2\pi); \\ [1, +\infty), & x = 2\pi. \end{cases}$$

5

Найти $\partial f(x)$, если $f(x) = |c_1^\top x| + |c_2^\top x|$

Решение: Пусть $f_1(x) = |c_1^\top x|$, а $f_2(x) = |c_2^\top x|$. Так как эти функции выпуклы, субдифференциал их суммы равен сумме субдифференциалов. Найдем каждый из них:

$$\partial f_1(x) = \partial \left(\max\{c_1^\top x, -c_1^\top x\} \right) = \begin{cases} -c_1, & c_1^\top x < 0 \\ \mathbf{conv}(-c_1; c_1), & c_1^\top x = 0 \\ c_1, & c_1^\top x > 0 \end{cases}$$

$$\partial f_2(x) = \partial \left(\max\{c_2^\top x, -c_2^\top x\} \right) = \begin{cases} -c_2, & c_2^\top x < 0 \\ \mathbf{conv}(-c_2; c_2), & c_2^\top x = 0 \\ c_2, & c_2^\top x > 0 \end{cases}$$

Далее интересными представляются лишь различные взаимные расположения векторов c_1 и c_2, рассмотрение которых предлагается читателю.

6

Найти $\partial f(x)$, если $f(x) = \|x\|_1$

Решение: По определению

$$\|x\|_1 = |x_1| + |x_2| + \ldots + |x_n| = s_1 x_1 + s_2 x_2 + \ldots + s_n x_n$$

Рассмотрим эту сумму как поточечный максимум линейных функций по x: $g(x) = s^\top x$, где $s_i = \{-1, 1\}$. Каждая такая функция однозначно определяется набором коэффициентов $\{s_i\}_{i=1}^n$.

Тогда по теореме Дубовицкого - Милютина, в каждой точке $\partial f = \mathbf{conv}\left(\bigcup\limits_{i \in I(x)} \partial g_i(x)\right)$

Заметим, что $\partial g(x) = \partial\left(\max\{s^\top x, -s^\top x\}\right) = \begin{cases} -s, & s^\top x < 0 \\ \mathbf{conv}(-s; s), & s^\top x = 0 \\ s, & s^\top x > 0 \end{cases}.$

Причем, правило выбора "активной" функции поточечного максимума в каждой точке следующее:

- Если j-ая координата точки отрицательна, $s_i^j = -1$
- Если j-ая координата точки положительна, $s_i^j = 1$
- Если j-ая координата точки равна нулю, то подходят оба варианта коэффициентов и соответствующих им функций, а значит, необходимо включать субградиенты этих функций в объединение в теореме Дубовицкого - Милютина.

В итоге получаем ответ:

$$\partial f(x) = \left\{g \ : \ \|g\|_\infty \leq 1, \quad g^\top x = \|x\|_1\right\}$$

References

- Lecture Notes for ORIE 6300: Mathematical Programming I by Damek Davis