МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ
им. М.В.Ломоносова

физический факультет

Кафедра компьютерных методов физики

А.И.ЧУЛИЧКОВ

ЭКСТРЕМАЛЬНЫЕ ЗАДАЧИ

КОНСПЕКТ ЛЕКЦИЙ

Москва, 1996

Конспект лекций спецкурса "Экстремальные задачи".

Экстремальные задачи формулируются как задачи отыскания максимального или минимального значения функции или функционала $J(u)$ на заданном множестве $\mathcal{U}$ некоторого пространства. Со значением функции $J(u)$ часто связывают цену (или качество) управления, а аргумент u задает само управление. Заметим, что задача максимизации функции $J(u)$ на множестве $\mathcal{U}$ эквивалентна задаче минимизации функции $-J(u)$ на том же множестве $\mathcal{U}$, поэтому можно ограничиться только задачами на минимум.

В курсе лекций рассматриваются задачи минимизации функций одного или нескольких переменных, определенных на некотором множестве $\mathcal{U}$ евклидова пространства (так называемые задачи без ограничений); задачи на минимакс для функций многих переменных, в которых требуется минимизировать функцию по одной группе переменных и максимизировать по другой, – геометрическим образом точек, в которых достигается минимакс, является поверхность типа седла; методы математического программирования, в которых множество, на котором минимизируется функция, задано в виде системы равенств и (или) неравенств. Рассматриваются также методы вариационного исчисления и оптимального управления.

Часть 1. Методы минимизации функций.

Глава 1. Минимизация функций одного переменного.

Общие положения. Постановка задачи. Пусть $\mathcal{R}_1 = \{u : -\infty < u < +\infty\}$ – числовая ось, $\mathcal{U} \subset \mathcal{R}_1$ – некоторое подмножество числовой оси, $J(\cdot)$ – числовая функция, заданная на $\mathcal{U}$ и принимающая в каждой точке $u \in \mathcal{U}$ конечные значения.

Определение. Точка $u_* \in \mathcal{U}$ называется точкой минимума функции $J(\cdot)$ на множестве $\mathcal{U}$, если $J(u_*) \le J(u)$ для всех $u \in \mathcal{U}$. Величину $J(u_*)$ назовем наименьшим, или минимальным, значением функции $J(\cdot)$ на множестве $\mathcal{U}$ и будем обозначать $\min\limits_{u\in\mathcal{U}} J(u) = J(u_*)$. Множество всех точек минимума обозначим $\mathcal{U}_*$.

В зависимости от свойств множества $\mathcal{U}$ и функции $J(\cdot)$ множество $\mathcal{U}_*$ может содержать одну или несколько точек, бесконечно много точек, или не содержать ни одной, то есть быть пустым.

Пример 1. Функция $J(u) = u^2$ на множестве $\mathcal{U} = \mathcal{R}_1$ имеет единственную точку минимума $\mathcal{U}_* = \{u = 0\}$; на множестве $\mathcal{U} = \{u : 1 < u \le 2\}$ множество $\mathcal{U}_*$ – пустое множество.

Пример 2. Функция $J(u) = u^4 - 2u^2$ на множестве $\mathcal{U} = \mathcal{R}_1$ имеет две точки минимума: $\mathcal{U}_* = \{u : u = \pm 1\}$.

Пример 3. Функция

$$J(u) = |u| + |u-1| - 1, \tag{1}$$

рассматриваемая на всей числовой прямой $\mathcal{U} = \mathcal{R}_1$, достигает своего минимального значения, равного нулю, во всех точках интервала $\mathcal{U}_* = \{u : 0 \le u \le 1\}$. Если же эту функцию рассматривать на интервале $\mathcal{U} = \{u : 1 \le u \le 2\}$, то множество $\mathcal{U}_*$ будет состоять из единственной точки $\mathcal{U}_* = \{u = 1\}$. Единственной точка минимума будет и для случая, когда $\mathcal{U} = \{u : 2 \le u \le 3\}$; при этом $\mathcal{U}_* = \{u = 2\}$. Наконец, если множество, на котором минимизируется функция (1), задано в виде $\mathcal{U} = \{u : 1 < u \le 2\}$, то $\mathcal{U}_* = \emptyset$.

Пример 4. Функция

$$J(u) = \begin{cases} \sin^2 \frac{\pi}{u}, & u \neq 0 \\ 0, & u = 0 \end{cases}, \tag{2}$$

рассматриваемая на множестве $\mathcal{U} = \{u : 1 \le u \le 2\}$, имеет единственную точку, в котором достигается ее минимальное значение; это точка $u_* = 1$. На интервале $\mathcal{U} = \{u : 1/3 \le u \le 1\}$ таких точек будет три, и $\mathcal{U}_* = \{u = 1/3, u = 1/2, u = 1\}$. Если же $\mathcal{U} = \{u : 0 < u \le 1\}$, то число точек минимума функции (2) уже бесконечно, и $\mathcal{U}_*$ – счетное множество, $\mathcal{U}_* = \{u : u = 1/k, k = 1, 2, \ldots$. Наконец, если $\mathcal{U} = \{u : 2 \le u < \infty\}$, то $\mathcal{U}_* = \emptyset$, так как для любого $b \in \mathcal{U}$ найдется такая точка $v \in \mathcal{U}$, что $v < u$ и $J(u) < J(v)$.

Пример 5. Функция

$$J(u) = \begin{cases} u, & u \neq 0 \\ 1, & u = 0 \end{cases},$$

и на множестве $\mathcal{U} = \{u : 0 \le u \le 1\}$, и на множестве $\mathcal{U} = \{u : 0 < u \le 1\}$, не имеет минимального значения; для нее $\mathcal{U}_* = \emptyset$.

Пример 6. Функция $J(u) = \ln u$ на своей области определения не имеет минимального значения, и для нее $\mathcal{U}_* = \emptyset$.

В примерах 1-4 мы имеем дело с непрерывными ограниченными снизу функциями, в примере 5 – с разрывной функцией, а в примере 6 – с неограниченной снизу функцией. Напомним эти определения.

Определение. Функция $J(\cdot)$ называется ограниченной снизу на $\mathcal{U}$, если существует такое число M, что $J(u) \ge M$ для любого $u \in \mathcal{U}$. Функция $J(\cdot)$ неограничена снизу на $\mathcal{U}$, если найдется такая последовательность $\{u_k\} \subset \mathcal{U}$, такая, что $\lim\limits_{k\to\infty} J(u_k) = -\infty$.

Если $\mathcal{U}_* = \emptyset$, естественным обобщением понятия минимального значения является понятие точной нижней грани.

Определение. Пусть $J(\cdot)$ ограничена снизу на множестве $\mathcal{U}$. Тогда число J_* называется точной нижней гранью функции $J(\cdot)$ на множестве $\mathcal{U}$, если выполнены следующие два условия:

1. $J_* \le J(u)$ для любого $u \in \mathcal{U}$;

2. для любого $\varepsilon > 0$ найдется точка $u_\varepsilon \in \mathcal{U}$, такая, что $J(u_\varepsilon) < J_* + \varepsilon$. Если функция $J(\cdot)$ неограничена снизу на множестве $\mathcal{U}$, то под точной нижней гранью понимается $J_* = -\infty$.

Точная нижняя грань обозначается следующим образом:

$$J_* = \inf_{u \in \mathcal{U}} J(u).$$

Если $\mathcal{U}_* \neq \emptyset$, то точная нижняя грань функции $J(\cdot)$ на множестве $\mathcal{U}$ существует и совпадает с минимальным значением функции $J(\cdot)$ на множестве $\mathcal{U}$, то есть

$$\min_{u \in \mathcal{U}} J(u) = \inf_{u \in \mathcal{U}} J(u).$$

В этом случае говорят, что функция $J(\cdot)$ достигает на $\mathcal{U}$ своей точной нижней грани. Заметим, что точная нижняя грань функции $J(\cdot)$ на множестве $\mathcal{U}$ всегда существует, в то время, как ее минимальное значение может быть неопределено.

Определение. Минимизирующей последовательностью для $J(\cdot)$ на множестве $\mathcal{U}$ назовем такую последовательность $\{u_k\} \subset \mathcal{U}$, для которой существует предел $\lim\limits_{k\to\infty} J(u_k)$, равный $J_* = \inf\limits_{u \in \mathcal{U}} J(u)$.

Из определения точной нижней грани следует, что минимизирующая последовательность всегда существует.

Определение. Последовательность $\{u_k\}$ сходится к некоторому множеству $\mathcal{U}$, если

$$\lim_{k\to\infty} \rho(u_k, \mathcal{U}) = 0,$$

где

$$\rho(u_k, \mathcal{U}) = \inf_{u \in \mathcal{U}} |u_k - u|$$

– расстояние от точки u_k до множества $\mathcal{U}$.

Если $\mathcal{U}_* \neq \emptyset$, то всегда существует минимизирующая последовательность, сходящаяся к $\mathcal{U}_*$. В качестве такой последовательности можно, например, взять стационарную последовательность $u_k = u_*$, $k = 1, 2, \ldots$, где u_* – некоторая точка из $\mathcal{U}_*$. Однако при $\mathcal{U}_* \neq \emptyset$ не всякая минимизирующая последовательность сходится к $\mathcal{U}_*$.

Пример 7. Рассмотрим функцию

$$J(u) = \begin{cases} \dfrac{1}{u^2}, & u \neq 0 \\ 0, & u = 0 \end{cases}.$$

Для нее минимизирующей является последовательность $u_k = k$, $k = 1, 2, \ldots$, – бесконечно большая, в то время как $U_* = \{u = 0\}$.

Можно привести пример, когда минимизирующая последовательность не сходится к множеству $\mathcal{U}_* \neq \emptyset$ и для непрерывной функции.

Пример 8. Для функции $J(u) = \dfrac{u^2}{1+u^4}$, заданной и непрерывной на всей числовой прямой, так же, как и в предыдущем примере, минимизирующей является последовательность $u_k = k$, $k = 1, 2, \ldots$, для которой $\lim u_k = \infty$, но однако $U_* = \{u = 0\}$.

В зависимости от целей, стоящих перед исследователем, занимающимся минимизацией функции, различают два типа задач. К первому типу относятся задачи, в которых требуется определить только значение точной нижней грани функции $J(\cdot)$, при этом не важно, является ли множество $\mathcal{U}_*$ пустым или не пустым. Ко второму типу относят те задачи, в которых $\mathcal{U}_* \neq \emptyset$, и требуется, наряду с определением значения J_*, указать хотя бы одно значение аргумента $u_* \in \mathcal{U}_*$, на котором достигается точная нижняя грань.

Точное решение задач как первого, так и второго типа возможно лишь для весьма неширокого класса задач. Поэтому для приближенного решения задачи первого типа обычно строят минимизирующую последовательность $\{u_k\} \in \mathcal{U}$, и в качестве приближенного значения J_* выбирают величину $J(u_k)$ с достаточно большим номером k. Для решения задач второго типа требуется еще и выбрать такую минимизирующую последовательность, которая сходилась бы к множеству $\mathcal{U}_*$ (если оно не пусто), и в качестве приближенного решения $(\tilde{J}_*, \tilde{u}_*)$ выбрать значения $(J(u_k), u_k)$ с достаточно большим номером k. Однако выяснить, сходится ли минимизирующая последовательность к $\mathcal{U}_*$, достаточно сложно, и выполнение этого условия требует применения специальных методов. Поэтому здесь пока будут рассмотрены лишь такие задачи второго типа, в которых любая минимизирующая последовательность сходится к $\mathcal{U}_*$. Один из классов таких задач выделяет так называемая теорема Вейерштрасса.

Теорема 1. Пусть $\mathcal{U}$ – замкнутое ограниченное множество из $\mathcal{R}_1$, $J(\cdot)$ – функция, непрерывная на $\mathcal{U}$. Тогда $J(\cdot)$ ограничена снизу на $\mathcal{U}$, множество $\mathcal{U}_*$ точек минимума $J(\cdot)$ на $\mathcal{U}$ непусто, замкнуто и любая минимизирующая последовательность сходится к $\mathcal{U}_*$.

Эта теорема позднее будет получена как следствие более общего утверждения.

Наряду с точкой минимума (который иногда называют глобальным минимумом) иногда интерес представляет нахождение точек локального минимума.

Определение. Точка $v_* \in \mathcal{U}$ называется точкой локального минимума функции $J(\cdot)$, если найдется такое число $\alpha > 0$, такое, что $J(v_*) \leq J(v)$ для всех $v \in \mathcal{U} \cap \{u : |u - v_*| < \alpha\} \equiv O_\alpha(v_*)$. Если при некотором α равенство $J(v_*) = J(v)$ для $v \in O_\alpha(v_*)$ возможно только при $v = v_*$, то v_* называется точкой строгого локального минимума.

Опишем класс функций, у которых точка локального минимума совпадает с точкой (глобального) минимума.

Определение. Функция $J(\cdot)$ называется унимодальной на отрезке $\mathcal{U} = \{u : a \leq u \leq b\}$, если она непрерывна на $[a, b]$ и найдутся такие числа α и β, $a \leq \alpha \leq \beta \leq b$, такие, что

1. $J(\cdot)$ строго монотонно убывает на отрезке $a \le u \le \alpha$ (если $a < \alpha$);
2. $J(\cdot)$ строго монотонно возрастает на отрезке $\beta \le u \le b$ (если $\beta < b$);
3. $J(u) = J_* = \inf\limits_{u \in \mathcal{U}} J(u)$ при $\alpha \le u \le \beta$.

При этом, очевидно, $\mathcal{U}_* = [\alpha, \beta]$.

Классический метод минимизации. Этот метод основан на результатах дифференциального исчисления. Пусть $J(\cdot)$ – функция, кусочно непрерывная и кусочно гладкая на отрезке $[a, b]$, что означает, что на $[a, b]$ может существовать лишь конечное число точек, в которых $J(\cdot)$ либо терпит разрыв первого рода, либо не имеет производной. Тогда, как известно, точками экстремума могут быть лишь те точки $u_* \in [a, b]$, в которых выполнено одно из следующих свойств:
1. либо $J(\cdot)$ терпит разрыв;
2. либо $J(\cdot)$ непрерывна и $J'(\cdot)$ не существует;
3. либо $J'(\cdot)$ существует и равна нулю;
4. либо $u_* = a$, либо $u_* = b$.

Точки, в которых выполнено одно из условий 1-4, называются точками, подозрительными на экстремум. Поиск минимума сводится к нахождению всех точек, подозрительных на экстремум, и выбору из них той точки, в которой достигается минимальное значение функции $J(\cdot)$, либо, если это требуется, выбору точки локального минимума.

Классический метод работает лишь в том случае, когда функция имеет достаточно простой вид. В более сложных случаях применяются численные методы минимизации.

Методы минимизации, как численные, так и точные, приводят к успеху лишь для достаточно гладких функций, минимизируемых на множествах достаточно простой структуры – например, на ограниченных замкнутых множествах. Если функция не является кусочно непрерывной или кусочно гладкой, а множество, на котором она минимизируется, не является ограниченным и замкнутым, то вряд ли может быть построен разумный способ ее минимизации. В частности, для функции $J(\cdot)$, не являющейся кусочно непрерывной на множестве $\mathcal{U} \in \mathcal{R}_1$, единственным способом поиска минимума может быть признан перебор всех точек множества $\mathcal{U}$, что в случае бесконечного множества $\mathcal{U}$ неприемлемо.

Численные методы минимизации. Золотое сечение. Опишем простейшие способы минимизации функций одного переменного, не требующие вычисления производных. Будем предполагать, что минимизируемая функция унимодальна на отрезке $[a, b]$.

Вычислим значения функции $J(u)$ на концах отрезка в точках $u = a$ и $u = b$, а также в двух внутренних точках u_1 и u_2, $a < u_1 < u_2 < b$, сравним полученные значения между собой и выберем из них наименьшее. Пусть для определенности это значение $J(u_1)$. Тогда, в соответствии с определением унимодальной функции, минимум расположен в одном из примыкающих к точке $u = u_1$ отрезков. Поэтому отрезок $[u_2, b]$ можно отбросить и минимизировать функцию на оставшемся отрезке $[a, u_2]$, для которого следует повторить ту же процедуру, то есть

выбрать две точки внутри него и вычислить значение функции $J(\cdot)$ в двух внутренних точках и на концах отрезка. Однако на предыдущем шаге мы уже нашли значения $J(\cdot)$ на концах этого отрезка и в одной внутренней точке – точке $u = u_1$, поэтому достаточно выбрать внутри отрезка $[a, u_2]$ еще одну точку u_3, вычислить значение функции $J(\cdot)$ в этой точке и провести сравнение полученных значений – это вчетверо уменьшает объем вычислений.

Как выбрать расположение внутренних точек получаемых отрезков? Мы всякий раз делим отрезок на три части, причем одна из точек деления уже определена предыдущими вычислениями. Если выбирать одну из частей отрезка слишком малой, то это выгодно, если она остается для дальнейшей процедуры, но не выгодно, если отбрасывается. Разумный компромисс состоит в выборе следующего отрезка так, чтобы он был подобен предыдущему. Это соображение приводит к соотношениям

$$b - u_2 = u_1 - a, \qquad \frac{b-a}{u_2-a} = \frac{b-u_1}{b-u_2},$$

откуда

$$\frac{b-a}{u_2-a} = \frac{u_2-a}{b-u_2},$$

то есть точка u_2 (и симметричная ей относительно середины отрезка $[a, b]$ точка u_1) делит весь отрезок $[a, b]$ в золотом отношении: длина большей части отрезка $[a, b]$ относится к длине меньшей его части так же, как длина всего отрезка $[a, b]$ относится к длине его большей части. Это отношение равно

$$\frac{u_2-a}{b-a} = \frac{\sqrt{5}-1}{2} = \varphi = 0.618...$$

В результате каждого шага остается отрезок, на котором производится поиск минимума функции $J(\cdot)$, длиной в $\varphi = 0.618...$ от длины предыдущего, откуда легко заключить, что метод деления отрезка в золотом отношении сходится, причем длина отрезка, на котором находится точка минимума функции $J(\cdot)$, убывает с числом шагов как член геометрической прогрессии со знаменателем $\varphi = 0.618....$ Итерации прекращаются, когда длина этого отрезка становится меньше требуемой точности определения точки минимума u_*. Заметим, однако, что точности определения величины минимума J_* можно оценить лишь в случае, когда $J(\cdot)$ является кусочно гладкой с ограниченной производной.

Этот метод можно применять и для минимизации функций, не являющихся унимодальными, однако при этом невозможно гарантировать, что полученное решение $J(u_n)$ даже при больших n будет близко к значению глобального минимума J_* функции $J(\cdot)$ на отрезке $[a, b]$. В таких случаях следует применять более изощренные методы построения минимизирующих последовательностей, один из которых мы рассмотрим.

Численные методы минимизации. Метод ломаных. Выделим класс функций, удовлетворяющих условию Липшица на отрезке $[a, b]$:

Определение. Функция $J(\cdot)$ удовлетворяет условию Липшица на отрезке $[a,b]$, если найдется число $L>0$, такое, что

$$|J(u)-J(v)|\le L|u-v| \quad \text{для любых} \quad u,v\in[a,b]. \tag{3}$$

Постоянная L называется константой Липшица функции $J(\cdot)$ на отрезке $[a,b]$.

Условие (3) имеет простой геометрический смысл: тангенс угла наклона $\dfrac{|J(u)-J(v)|}{|u-v|}$ хорды, соединяющей точки $(u,J(u))$ и $(v,J(v))$ графика функции $J(\cdot)$ на отрезке $[a,b]$ не превышает константы L для всех u, $v\in[a,b]$.

Из условия (3) следует, что функция $J(\cdot)$ непрерывна на отрезке $[a,b]$, и, согласно теореме Вейерштрасса, множество $\mathcal{U}_*$ точек минимума функции $J(\cdot)$ на отрезке $[a,b]$ непусто.

Для построения метода ломаных необходимо знать константу Липшица для функции $J(\cdot)$ на отрезке $[a,b]$. Приведем два простых утверждения, позволяющих оценить эту константу.

Утверждение 1. Пусть функция $J(\cdot)$ непрерывна на отрезке $[a,b]$ и на каждом отрезке $[a_i,a_{i+1}]$, $i=1,2,\dots,n$, $a=a_1\le a_2\le\ldots\le a_{n+1}=b$, удовлетворено условие Липшица с константой L_i; тогда $J(\cdot)$ на $[a,b]$ удовлетворяет условию Липшица с константой $L=\max\limits_{1\le i\le n}L_i$.

Утверждение 2. Пусть функция $J(\cdot)$ дифференцируема на отрезке $[a,b]$ и ее производная $J'(\cdot)$ ограничена на этом отрезке. Тогда $J(\cdot)$ на $[a,b]$ удовлетворяет условию Липшица с константой $L=\sup\limits_{u\in[a,b]}|J'(u)|$.

Идея метода ломаных состоит в том, что функция $J(\cdot)$ на отрезке $[a,b]$ аппроксимируется кусочно линейной функцией, и вместо минимума $J(\cdot)$ на отрезке $[a,b]$ находится минимум кусочно линейной функции. Прежде, чем приступить к описанию этого метода, заметим, что функция $J(\cdot)$, удовлетворяющая на $[a,b]$ условию Липшица (3) с константой L, обладает следующими свойствами.

Пусть $J(\cdot)$ на отрезке $[a,b]$ удовлетворяет условию Липшица (3) с константой L. Зафиксируем некоторую точку $v\in[a,b]$ и построим функцию

$$g(u,v)=J(v)-L|u-v| \tag{4}$$

аргумента $u\in[a,b]$. Функция $g(u,v)$ как функция аргумента u является кусочно линейной на $[a,b]$, и график ее представляет собой два луча с коэффициентами $+L$ и $-L$, сходящиеся в одной вершине - точке с координатами $(v,J(v))$. Кроме того, в силу соотношений (3) и (4), выполнено неравенство

$$J(u)-g(u,v)=J(u)-J(v)+L|u-v|\ge L|u-v|-|J(u)-J(v)|\ge 0,$$

то есть

$$J(u)\ge g(u,v) \quad \text{для всех} \quad u\in[a,b],$$

причем $J(v) = g(v, v)$, то есть график функции $J(\cdot)$ на отрезке $[a, b]$ лежит не ниже ломаной $g(u, v)$ и имеет с ней общую точку $(v, J(v))$.

Построим теперь последовательность кусочно линейных функций, аппроксимирующую функцию $J(\cdot)$ на отрезке $[a, b]$, и выберем соответствующую минимизирующую последовательность. Для этого на первом шаге выберем произвольную точку $u_0 \in [a, b]$ и построим функцию $g(u, u_0) = J(u_0) - L|u - u_0| = p_0(u)$. Следующую точку u_1 определим из условия

$$p_0(u_1) = \min_{u \in [a,b]} p_0(u), \qquad u_1 \in [a, b],$$

причем очевидно, что либо $u_1 = a$, либо $u_1 = b$. Далее строится новая функция

$$p_1(u) = \max\{g(u, u_1), p_0(u)\},$$

и очередная точка u_2 находится как решение задачи на минимум

$$p_1(u_2) = \min_{u \in [a,b]} p_1(u), \qquad u_2 \in [a, b].$$

Далее процедура повторяется. Пусть выбраны точки $u_0, u_1, \ldots, u_n$, $n \geq 1$. Тогда составляется функция

$$p_n(u) = \max\{g(u, u_n), p_{n-1}(u)\} = \max_{1 \leq i \leq n} \{g(u, u_i)\},$$

и следующая точка u_{n+1} выбирается из условия

$$p_n(u_{n+1}) = \min_{u \in [a,b]} p_n(u), \qquad u_{n+1} \in [a, b].$$

Если минимум в последнем соотношении достигается в нескольких точках, то в качестве u_{n+1} выбирается любая из них.

Функция $p_i(\cdot)$ является кусочно линейной на $[a, b]$, ее график состоит из отрезков прямых с угловыми коэффициентами $\pm L$, проходит не выше графика функции $J(\cdot)$ на отрезке $[a, b]$ и для всех $u \in [a, b]$ выполнено неравенство

$$p_{n-1}(u) = \max_{1 \leq i \leq n-1} \{g(u, u_i)\} \leq \max_{1 \leq i \leq n} \{g(u, u_i)\} = p_n(u)$$

и, кроме того,

$$p_n(u) \leq J(u), \qquad u \in [a, b],$$

то есть в каждой точке $u \in [a, b]$ последовательность $\{p_n(u)\}$ не убывает и ограничена сверху значением $J(u)$.

Таким образом, на каждом шаге метода ломаных минимизация функции $J(\cdot)$ на отрезке $[a, b]$ заменяется более простой задачей минимизации кусочно линейной функции $p_n(\cdot)$.

Сходимость и точность метода ломаных даются следующей теоремой.

Теорема 2. Пусть $J(\cdot)$ – произвольная функция, удовлетворяющая на отрезке $[a,b]$ условию (3).Тогда последовательность $\{u_n\}$, полученная с помощью описанного выше метода ломаных, обладает следующими свойствами:

1.

$$\lim_{n\to\infty} J(u_n) = \lim_{n\to\infty} p_n(u_{n+1}) = J_* = \inf_{u\in[a,b]} J(u),$$

причем справедлива оценка

$$0 \le J(u_{n+1}) - J_* \le J(u_{n+1}) - p_n(u_{n+1}), \quad n = 0,1,2,\ldots$$

2. Последовательность $\{u_n\}$ сходится к множеству $\mathcal{U}_*$ точек минимума функции $J(\cdot)$ на отрезке $[a,b]$.

Доказательство. Возьмем произвольную точку $u_* \in \mathcal{U}_*$. По построению функции $p_n(\cdot)$,

$$p_{n-1}(u_n) = \min_{u\in[a,b]} p_{n-1}(u) \le p_{n-1}(u_{n+1}) \le p_n(u_{n+1}) =$$

$$= \min_{u\in[a,b]} p_n(u) \le p_n(u_*) \le J(u_*) = J_*,$$

то есть последовательность $\{p_n(u_{n+1})\}$ монотонно не убывает и ограничена сверху значением $J(u_*)$. Следовательно, существует предел этой последовательности, причем

$$\lim_{n\to\infty} p_n(u_{n+1}) = p_* \le J_*. \tag{5}$$

Покажем, что $p_* = J_*$. Так как отрезок $[a,b]$ является ограниченным множеством, то из последовательности $\{u_n\} \subset [a,b]$ можно выбрать подпоследовательность $\{u_{n_k}\}$, сходящуюся к некоторой точке $v_* \in [a,b]$. Так как для всех $i = 1,2,\ldots,n$ выполняется равенство $p_n(u_i) = J(u_i)$, то при любом n и при $i = 1,2,\ldots,n$ можно записать

$$0 \le p_n(u_i) - \min_{u\in[a,b]} p_n(u) = J(u_i) - p_n(u_{n+1}) = p_n(u_i) - p_n(u_{n+1}) \le L|u_i - u_{n+1}|.$$

Принимая здесь $n = n_k - 1$, $i = n_{k-1} \le n_k - 1$, получим

$$0 \le J(u_{n_{k-1}}) - p_{n_k-1}(u_{n_{k-1}}) \le L|u_{n_{k-1}} - u_{n_k}|, \quad k \ge 2.$$

Устремим в последних неравенствах $k \to \infty$, и в силу сходимости подпоследовательности $\{u_{n_k}\}$ к точке v_*, учитывая соотношение (5), получим равенства $\lim\limits_{k\to\infty} p_{n_k-1}(u_{n_k}) = \lim\limits_{k\to\infty} J(u_{n_k}) = p_* = J_*$. Так как v_* – произвольная предельная точка последовательности $\{u_n\}$, то первое утверждение теоремы доказано. Справедливость второго утверждения следует из теоремы 1.

Выпуклые функции. Более эффективные методы минимизации могут быть предложены для специального класса функций, называемых выпуклыми.

Определение. Функция $J(\cdot)$, определенная на отрезка $[a,b]$, называется выпуклой на $[a,b]$, если

$$J(\alpha u+(1-\alpha)v) \le \alpha J(u)+(1-\alpha)J(v) \quad \text{для всех } u,v \in [a,b] \quad \text{и для всех} \quad \alpha \in [0,1]. \tag{6}$$

Соотношению (6) можно дать следующую геометрическую интерпретацию: когда α пробегает все множество $[0,1]$, точки $(\alpha u + (1-\alpha)v, \alpha J(u) + (1-\alpha)J(v))$ на плоскости переменных (u,J) пробегают хорду AB, соединяющую точку A с координатами $(u,J(u))$, с точкой $B = (v,J(v))$ на графике функции $J = J(u)$, $u \in [a,b]$. Неравенство (6) означает, что эта хорда всегда лежит не ниже графика функции $J(\cdot)$ на $[a,b]$.

Приведем несколько утверждений относительно свойств выпуклых функций.

Утверждение 3. Для выпуклости функции $J(\cdot)$ на $[a,b]$ необходимо и достаточно, чтобы для всех точек u, v и w, таких, что $a \le v < u < w \le b$ выполнялось неравенство

$$\frac{J(u)-J(v)}{u-v} \le \frac{J(w)-J(v)}{w-v} \le \frac{J(w)-J(u)}{w-u}. \tag{7}$$

Геометрический смысл соотношения (7) становится ясен, если заметить, что $\dfrac{J(w)-J(u)}{w-u}$ есть угловой коэффициент хорды, соединяющей точки $(w,J(w))$ и $(u,J(u))$. Т.о., в частности, угол наклона хорды, соединяющей две точки $(u,J(u))$ и $(u+h,J(u+h))$ графика выпуклой функции $J(\cdot)$ к оси OX неубывает по u (разностная производная выпуклой функции — неубывает).

Доказательство. Пусть $J(\cdot)$ — выпуклая на $[a,b]$ функция, тогда для $v,w \in [a,b]$ для любого $u \in [v,w]$ выполнено $u = \alpha v + (1-\alpha)w \in [a,b]$ при

$$\alpha = \frac{w-u}{w-v}.$$

Тогда

$$J(u) = J(\alpha v + (1-\alpha)w) \le \alpha J(v) + (1-\alpha)J(w) = \frac{w-u}{w-v}J(v) + \frac{u-v}{w-v}J(w);$$

после преобразований получим

$$(w-v)(J(u)-J(v)) \le (u-v)(J(w)-J(v)),$$

или

$$(w-u)(J(w)-J(v)) \le (w-v)(J(w)-J(u)).$$

Пусть теперь выполнено неравенство (7). Проводя рассуждения в обратном порядке, получим выпуклость функции $J(\cdot)$.

Теорема 3. Выпуклая на $[a,b]$ функция $J(\cdot)$ в каждой внутренней точке u отрезка $[a,b]$ непрерывна и имеет конечную правую производную

$$J'(u+0) = \lim_{h \to +0} \frac{J(u+h)-J(u)}{h}$$

и конечную левую производную

$$J'(u-0) = \lim_{h \to +0} \frac{J(u) - J(u-h)}{h},$$

причем $J'(u-0) \le J'(u+0)$ при всех $u \in (a,b)$.

Доказательство. Согласно утверждению 3, при всех τ и h, таких, что $0 < h < \tau$ и при которых точки u, $u \pm h$, $u \pm \tau$ принадлежат интервалу (a,b), выполняются неравенства

$$\frac{J(u) - J(u-\tau)}{\tau} \le \frac{J(u) - J(u-h)}{h} \le \frac{J(u+h) - J(u)}{h} \le \frac{J(u+\tau) - J(u)}{\tau} \qquad (8)$$

которые означают, что величина $\dfrac{J(u+h) - J(u)}{h}$ монотонно убывает при убывании h и ограничена снизу, например, величиной $\dfrac{J(u) - J(u-\tau)}{\tau}$, не зависящей от h. Отсюда следует существование правой производной $J'(u+0)$. Аналогично доказывается и существование левой производной. Неравенства (8) позволяют заключить также, что выполняется и соотношение $J'(u-0) \le J'(u+0)$ между односторонними производными. Из существования правой и левой производных во внутренних точках отрезка $[a,b]$ следует и непрерывность $J(\cdot)$ на интервале (a,b).

Заметим, что выпуклая функция может не иметь конечных односторонних производных на концах отрезка $[a,b]$, и значит, может терпеть разрыв в точках $u = a$ или (и) $u = b$.

Пример 9. Функция

$$J(u) = \begin{cases} u, & 0 < u < 1 \\ 2, & u = 0 \ u = 1 \end{cases}$$

является выпуклой на отрезке $[a,b]$, причем на концах отрезка, в точках $u = 0$ и $u = 1$, терпит разрыв и не имеет соответствующих односторонних производных.

Пример 10. Функция $J(u) = -\sqrt{1-u^2}$ выпукла и непрерывна на отрезке $[-1,1]$, но на концах его не имеет конечных односторонних производных $J'(-1+0)$ и $J'(1-0)$.

Утверждение 4. Пусть $J(\cdot)$ выпукла на $[a,b]$ и имеет конечные производные $J'(a+0)$ и $J'(b-0)$. Тогда неравенства

$$J'(a+0)(u-v) \le J(u) - J(v) \le J'(b-0)(u-v)$$

выполнены при всех u и v, таких, что $a \le v \le u \le b$, так, что $J(\cdot)$ удовлетворяет условию Липшица (3) с константой $L = \max\{|J'(a+0)|, |J'(b-0)|\}$.

Это утверждение показывает, что существование односторонних производных $J'(a+0)$ и $J'(b-0)$ на концах отрезка $[a,b]$ важно для выполнения условия Липшица.

Утверждение 5. Пусть $J(\cdot)$ выпукла на $[a,b]$. Тогда производные $J'(u+0)$ и $J'(u-0)$ монотонно возрастают на (a,b), а если существуют и производные $J'(a+0)$ и $J'(b-0)$, то это свойство выполняется и на всем отрезке $[a,b]$.

Утверждение 6. Если функция $J(\cdot)$ выпукла на $[a,b]$, $\lim\limits_{u\to a+0} J(u) = J(a)$ и $\lim\limits_{u\to b-0} J(u) = J(b)$, то $J(\cdot)$ унимодальна на $[a,b]$.

Утверждение 7. Для того, чтобы дифференцируемая на отрезке $[a,b]$ функция $J(\cdot)$ была выпукла на $[a,b]$, необходимо и достаточно, чтобы ее производная не убывала на $[a,b]$.

Утверждение 8. Для того, чтобы дважды дифференцируемая на отрезке $[a,b]$ функция $J(\cdot)$ была выпукла на $[a,b]$, необходимо и достаточно, чтобы $J''(u) \geq 0$ для всех $u \in [a,b]$.

Теорема 4. Пусть функция $J(\cdot)$ выпукла на $[a,b]$ и $\lim\limits_{u\to a+0} J(u) = J(a)$ и $\lim\limits_{u\to b-0} J(u) = J(b)$. Тогда множество $\mathcal{U}_*$ точек ее глобального минимума на $[a,b]$ непусто и все точки локального минимума функции $J(\cdot)$ на $[a,b]$ принадлежат множеству $\mathcal{U}_*$. Для того, чтобы $u_* \in \mathcal{U}_*$, необходимо и достаточно, чтобы в точке u_* выполнялись неравенства

$$J'(u_*+0) \geq 0, \qquad J'(u_*-0) \leq 0. \tag{9}$$

Если $u_* = a$ или $u_* = b$, то (9) заменяется одним неравенством $J'(a+0) \geq 0$ или $J'(b-0) \leq 0$ соответственно.

Доказательство.

Необходимость. Из условия на функцию $J(\cdot)$ и теоремы 3 следует непрерывность $J(\cdot)$ на $[a,b]$, а значит, по теореме 1 Вейерштрасса, $U_* \neq \emptyset$. Пусть $u_* \in [a,b]$ – некоторая точка локального минимума функции $J(\cdot)$. Тогда существует некоторая достаточно малая окрестность $O_\alpha(u_*)$ точки u_*, такая, что $J(u_*+h)-J(u) \geq 0$ для всех точек $u_* + h \in O_\alpha(u_*)$. Разделив это неравенство на $h > 0$ и на $h < 0$, устремляя h к нулю получим условия (9). Существование и конечность производных $J'(u \pm 0)$ следует из теоремы 3. Аналогично доказывается и необходимость других утверждений теоремы.

Достаточность. Пусть теперь некоторая точка $u \in (a,b)$ удовлетворяет условиям (9). В силу выпуклости функции $J(\cdot)$ на $[a,b]$ для всех $u \in [a,b]$ и всех $\alpha \in [0,1]$ выполнено неравенство

$$J(\alpha u + (1-\alpha)u_*) \leq \alpha J(u) + (1-\alpha)J(u_*),$$

откуда

$$J(u_* + \alpha(u-u_*)) - J(u_*) \leq \alpha(J(u) - J(u_*)).$$

Поделив обе части этого неравенства на α и устремив $\alpha \to +0$, получим

$$0 \leq J'(u_*+0)(u-u_*) \leq J(u) - J(u_*) \text{ при } u > u_*,$$

и

$$0 \le J'(u_* - 0)(u - u_*) \le J(u) - J(u_*) \text{ при } u < u_*,$$

где учтено (9), а эти неравенства и означают, что $J(u) \ge J(u_*)$ для всех $u \in [a, b]$, то есть $u_* \in \mathcal{U}_*$. Отсюда также следует, что всякая точка локального минимума является и точкой глобального минимума.

Численные методы минимизации выпуклых функций. Пусть $J(\cdot)$ – дифференцируемая и выпуклая на отрезке $[a, b]$ функция, тогда она удовлетворяет условию Липшица и унимодальна на $[a, b]$. Значит, для минимизации $J(\cdot)$ можно использовать описанный выше метод ломаных. Однако, основываясь на выпуклости $J(\cdot)$, предложим более эффективный способ минимизации – метод касательных, в котором в качестве отрезков прямых используют касательные к графику функции $J(\cdot)$ на отрезке $[a, b]$.

Действуя по аналогии с методом ломаных, зафиксируем некоторую точку $v \in [a, b]$ и построим функцию

$$g(u, v) = J(v) - J'(v)(u - v)$$

аргумента $u \in [a, b]$. Тогда, по свойствам выпуклых функций, выполнено неравенство

$$J(u) \ge g(u, v) \quad \text{для всех} \quad u \in [a, b],$$

причем $J(v) = g(v, v)$, то есть график функции $J(\cdot)$ на отрезке $[a, b]$ лежит не ниже касательной $g(u, v)$ к $J(\cdot)$ в точке v и имеет с ней единственную общую точку $(v, J(v))$.

Далее так же, как и в методе ломаных, построим последовательность кусочно линейных функций, аппроксимирующую функцию $J(\cdot)$ на отрезке $[a, b]$, и выберем соответствующую минимизирующую последовательность. Для этого на первом шаге выберем произвольную точку $u_0 \in [a, b]$, построим функцию $p_0(u) = g(u, u_0) = J(u_0) - J'(v)(u - v)$ и определим следующую точку u_1 из условия

$$p_0(u_1) = \min_{u \in [a,b]} p_0(u), \qquad u_1 \in [a, b],$$

причем очевидно, что при $J'(u_0) \ne 0$ либо $u_1 = a$, либо $u_1 = b$. Далее строится новая функция

$$p_1(u) = \max\{g(u, u_1), p_0(u)\},$$

и очередная точка u_2 находится как решение задачи на минимум

$$p_1(u_2) = \min_{u \in [a,b]} p_1(u), \qquad u_2 \in [a, b].$$

Далее процедура повторяется точно так же, как и в методе ломаных: пусть выбраны точки $u_0, u_1, \ldots, u_n$, $n \ge 1$. Тогда составляется функция

$$p_n(u) = \max\{g(u, u_n), p_{n-1}(u)\} = \max_{1 \le i \le n}\{g(u, u_i)\},$$

и следующая точка u_{n+1} выбирается из условия

$$p_n(u_{n+1}) = \min_{u\in[a,b]} p_n(u), \qquad u_{n+1} \in [a,b].$$

Функция $p_i(\cdot)$ является кусочно линейной на $[a,b]$, ее график состоит из отрезков касательных к функции $J(\cdot)$, проходит не выше графика функции $J(\cdot)$ на отрезке $[a,b]$ и для всех $u \in [a,b]$ последовательность $\{p_n(u)\}$ не убывает и ограничена сверху значением $J(u)$, то есть последовательность функций$\{p_n(u)\}$ сходится поточечно.

Заметим, что если на k-том шаге $J'(u_k) = 0$, то в точке u_k достигается минимум функции $J(\cdot)$ на $[a,b]$. В общем случае сходимость и точность метода касательных даются следующей теоремой.

Теорема 5. Пусть $J(\cdot)$ – выпуклая и дифференцируемая на отрезке $[a,b]$ функция. Тогда последовательность $\{u_n\}$, полученная с помощью описанного выше метода касательных, обладает следующими свойствами:

1.

$$\lim_{n\to\infty} J(u_n) = \lim_{n\to\infty} p_n(u_{n+1}) = J_* = \inf_{u\in[a,b]} J(u),$$

причем справедлива оценка

$$0 \le J(u_{n+1}) - J_* \le J(u_{n+1}) - p_n(u_{n+1}), \quad n = 0,1,2,\ldots$$

2. Последовательность $\{u_n\}$ сходится к множеству $\mathcal{U}_*$ точек минимума функции $J(\cdot)$ на отрезке $[a,b]$ и имеет не более двух предельных точек, совпадающих либо с точкой $u_* = \inf \mathcal{U}_*$, либо с точкой $u^* = \sup \mathcal{U}_*$.

Доказательство теоремы 5 аналогично доказательству теоремы 2.

Глава 2. Минимизация функций в конечномерном пространстве.

Пусть $\mathcal{R}_n$ – евклидово n-мерное пространство, $n < \infty$. Рассмотрим задачу минимизации функции $J(\cdot)$, определенной на некотором подмножестве $\mathcal{U}$ евклидова пространства $\mathcal{R}_n$, которое, возможно, совпадает со всем пространством $\mathcal{R}_n$: $\mathcal{U} \subseteq \mathcal{R}_n$. Будем рассматривать лишь такие функции $J(\cdot)$, которые принимают конечные значения в каждой точке множества $\mathcal{U}$.

Точно так же, как и для функций одного переменного, можно определить понятия точки минимума, точной нижней грани функции $J(\cdot)$ на $\mathcal{U}$, минимизирующей последовательности, локального минимума, сходимости последовательности точек из $\mathcal{R}_n$ к множеству $\mathcal{U} \subseteq \mathcal{R}_n$, только под точкой $u \in \mathcal{U}$ теперь понимается элемент n-мерного евклидова пространства, который в некотором ортонормированном базисе $\mathcal{R}_n$ может быть представлен как набор из n чисел – своих координат в этом базисе. Расстояние между точками теперь понимается как норма разности элементов u и v:

$$\rho(u, v) = \|u - v\| = \sqrt{(u - v, u - v)},$$

где $(\cdot, \cdot)$ – скалярное произведение в $\mathcal{R}_n$. Если $u, v \in \mathcal{U}$ заданы своими координатами $u = (u_1, u_2, \ldots, u_n)$ и $v = (v_1, v_2, \ldots, v_n)$ в некотором ортонормированном базисе $\mathcal{R}_n$, то

$$(u, v) = \sum_{i=1}^{n} u_i v_i.$$

Под ε–окрестностью $O_\varepsilon(u)$ точки $u \in \mathcal{U}$ будем понимать открытый шар в $\mathcal{R}_n$ с центром в точке u и радиусом, равным ε: $O_\varepsilon(u) = \{\tilde{u} \in \mathcal{R}_n : \|u - \tilde{u}\| < \varepsilon\}$. Точка u_0 называется предельной точкой множества $\mathcal{U}$, если любая ее ε-окрестность содержит точки множества $\mathcal{U}$, отличные от u.

Множество точек, в которых функция $J(\cdot)$ на $\mathcal{U}$ принимает минимальные значения, будем обозначать $\mathcal{U}_*$. Точно так же, как и для функций одного переменного, рассмотренных в первой главе, будем различать задачи на минимум двух типов. В задачах первого типа требуется определить точную нижнюю грань $J_* = \inf\limits_{u \in \mathcal{U}} J(u)$ функции $J(\cdot)$ на множестве $\mathcal{U}$, при этом не интересуясь, существуют ли такие точки множества $\mathcal{U}$, в которых эта точная нижняя грань достигается; таким образом, для задач первого типа не важно, является ли множество $\mathcal{U}_*$ пустым или нет.

В задачах второго типа, помимо отыскания точной нижней грани $J_* = \inf\limits_{u \in \mathcal{U}} J(u)$ функции $J(\cdot)$ на множестве $\mathcal{U}$, требуется указать хотя бы одну точку $u_* \in \mathcal{U}_* = \{u : u \in \mathcal{U}, J(u) = J_*\}$, либо доказать, что $\mathcal{U}_* = \emptyset$.

Далее нам понадобятся следующие определения.

Определение. Множество $\mathcal{U} \in \mathcal{R}_n$ называется компактным, если для любой последовательности $\{u_k\} \subset \mathcal{U}$ можно выделить подпоследовательность, сходящуюся к точке $u_0 \in \mathcal{U}$.

Согласно теореме Больцано-Вейерштрасса, гласящей, что всякая ограниченная последовательность имеет хотя бы одну предельную точку, в $\mathcal{R}_n$ компактными являются ограниченные замкнутые множества, и в конечномерном евклидовом пространстве ими и исчерпываются компактные множества.

Определение. Число a называется нижним [верхним] пределом ограниченной снизу [сверху] числовой последовательности $\{x_k\}$ и обозначается $\liminf\limits_{k\to\infty} x_k = a$ $[\limsup\limits_{k\to\infty} x_k = a]$, если

1. существует хотя бы одна подпоследовательность $\{x_{k_m}\}$ последовательности $\{x_k\}$, сходящаяся к a;
2. все предельные точки $\{x_k\}$ не меньше [не больше] числа a.

Иными словами, число a является наименьшей [наибольшей] предельной точкой последовательности $\{x_k\}$. В том случае, когда $\{x_k\}$ неограничена снизу [сверху], $\liminf\limits_{k\to\infty} x_k = -\infty$ $[\limsup\limits_{k\to\infty} x_k = \infty]$. Если $\lim\limits_{k\to\infty} x_k = -\infty$, то принимают $\limsup\limits_{k\to\infty} x_k = -\infty$, если же $\lim\limits_{k\to\infty} x_k = +\infty$, то принимают $\liminf\limits_{k\to\infty} x_k = +\infty$.

Определение. Функция $J(\cdot)$, определенная на $\mathcal{U} \in \mathcal{R}_n$, называется полунепрерывной снизу [сверху] в точке $u \in \mathcal{U}$, если для любой последовательности $\{u_k\} \subset \mathcal{U}$, сходящейся к точке u, имеет место неравенство $\liminf\limits_{k\to\infty} J(u_k) \geq J(u)$ $[\limsup\limits_{k\to\infty} J(u_k) \leq J(u)]$. Функция $J(\cdot)$, полунепрерывна снизу [сверху] на всем множестве $\mathcal{U}$, если она полунепрерывна снизу [сверху] в каждой точке этого множества.

Утверждение 1. Функция $J(\cdot)$ полунепрерывна снизу [сверху] в точке $u_0 \in \mathcal{U}$ тогда и только тогда, когда для любого $\varepsilon > 0$ найдется такое число $\delta > 0$, такое, что для всех $u \in \mathcal{U}$, для которых выполнено неравенство $\|u - u_0\| < \delta$, выполняется и $J(u) \geq J(u_0) - \varepsilon$ $[J(u) \leq J(u_0) + \varepsilon]$.

Утверждение 2. Функция непрерывна тогда и только тогда, когда она полунепрерывна снизу и полунепрерывна сверху.

Пример 1. Пусть $\mathcal{U} = \{u \in \mathcal{R}_n, \|u\| \leq 1\}$, а

$$J(u) = \begin{cases} \|u\|, & u \neq 0 \\ a, & u = 0 \end{cases}.$$

Тогда при $a \leq 0$ функция $J(\cdot)$ полунепрерывна снизу, а при $a \geq 0$ функция $J(\cdot)$ полунепрерывна сверху. При $a = 0$ функция $J(\cdot)$ непрерывна.

Пример 2. Пусть $\mathcal{U} = [-1, 1] \subset \mathcal{R}_1$,

$$J(u) = \begin{cases} u, & 0 < u \leq 1 \\ 1 - u, & -1 \leq u < 0 \\ a, & u = 0 \end{cases}.$$

Тогда при $a \leq 0$ функция $J(\cdot)$ полунепрерывна снизу, при $a \geq 1$ функция $J(\cdot)$ полунепрерывна сверху, а при $0 < a < 1$ функция $J(\cdot)$ не является ни полунепрерывной снизу, ни полунепрерывной сверху.

Определение. Множество

$$M(c) = \{u \in \mathcal{U}, \quad J(u) \leq c\}$$

точек из $\mathcal{U}$, таких, в которых значение функции $J(\cdot)$ не превосходит константу c, называется множеством Лебега функции $J(\cdot)$ на $\mathcal{U}$.

Следующее утверждение дает связь между полунепрерывностью снизу функции $J(\cdot)$ и замкнутостью ее множеств Лебега.

Утверждение 3. Пусть множество $\mathcal{U} \subseteq \mathcal{R}_n$ замкнуто. Тогда для того, чтобы функция $J(\cdot)$ была полунепрерывна снизу на $\mathcal{U}$, необходимо и достаточно, чтобы ее множества Лебега $M(c)$ были замкнуты при всех $c \in \mathcal{R}_1$[1]. В частности, если $J(\cdot)$ – полунепрерывна снизу, то множество $\mathcal{U}_*$ точек минимума $J(\cdot)$ на $\mathcal{U}$ замкнуто.

Сформулируем теперь теорему Вейерштрасса.

Теорема 1. Пусть $\mathcal{U} \subseteq \mathcal{R}_n$ – компактное множество, а функция $J(\cdot)$ определена, конечна и полунепрерывна снизу на множестве $\mathcal{U}$. Тогда $J_* = \inf\limits_{u \in \mathcal{U}} J(u) > -\infty$, множество $\mathcal{U}_* = \{u \in \mathcal{U}, J(u) = J_*\}$ непусто, компактно, и любая минимизирующая последовательность сходится к $\mathcal{U}_*$.

Замечание. Условие компактности в теореме 1 может быть заменено менее ограничительным, но при этом на функцию $J(\cdot)$ накладываются несколько более жесткие требования. В частности, справедливы следующие две теоремы.

Теорема 1^*. Пусть $\mathcal{U} \subseteq \mathcal{R}_n$ – непустое замкнутое подмножество $\mathcal{R}_n$, а функция $J(\cdot)$ определена, конечна и полунепрерывна снизу на множестве $\mathcal{U}$, и для некоторой точки $v \in \mathcal{U}$ множество Лебега $M(J(v)) = \{u \in \mathcal{U}, \quad J(u) \leq J(v)\}$ ограничено. Тогда $J_* = \inf\limits_{u \in \mathcal{U}} J(u) > -\infty$, множество $\mathcal{U}_* = \{u \in \mathcal{U}, J(u) = J_*\}$ непусто, компактно, и любая минимизирующая последовательность, принадлежащая $M(J(v))$, сходится к $\mathcal{U}_*$.

Теорема 1^{**}. Пусть $\mathcal{U} \subseteq \mathcal{R}_n$ – непустое замкнутое подмножество $\mathcal{R}_n$, а функция $J(\cdot)$ определена, конечна и полунепрерывна снизу на множестве $\mathcal{U}$, и для любой последовательности $\{u_k\} \subset \mathcal{U}$, для которой $\lim\limits_{k \to \infty} \|u_k\| = +\infty$ (если такие существуют), выполнено соотношение $\lim\limits_{k \to \infty} J(u_k) = +\infty$. Тогда $J_* = \inf\limits_{u \in \mathcal{U}} J(u) > -\infty$, множество $\mathcal{U}_* = \{u \in \mathcal{U}, J(u) = J_*\}$ непусто, компактно, и любая минимизирующая последовательность сходится к $\mathcal{U}_*$.

Заметим, что если в условиях теоремы 1^{**} не существует таких последовательностей, что $\lim\limits_{k \to \infty} \|u_k\| = +\infty$, то множество $\mathcal{U}$ ограничено, а значит – и компактно, и утверждение теоремы 1^{**} следует из теоремы 1 (Вейерштрасса).

Классический метод минимизации функций в конечномерных евклидовых пространствах. Этот метод эффективен, когда функция $J(\cdot)$ задана на всем евклидовом пространстве $\mathcal{R}_n$ и дважды дифференцируема на нем.

[1]Напомним, что пустое множество является замкнутым по определению.

Напомним, что функция $J(\cdot)$ называется дифференцируемой в точке $u \in \mathcal{R}_n$, если она определена в некоторой окрестности $O_\varepsilon(u)$ этой точки и существует вектор $J'(u) \in \mathcal{R}_n$ такой, что для любой точки $h \in \mathcal{R}_n$, для которой $u+h \in O_\varepsilon(u)$, функция $J(\cdot)$ представима в виде

$$J(u+h) = J(u) + (J'(u), h) + o(h, u),$$

где $(\cdot, \cdot)$ – скалярное произведение в $\mathcal{R}_n$, а $o(h, u)$ – бесконечно малая более высокого порядка, чем $\|h\|$ при $h \to 0$, то есть $\lim\limits_{h\to 0} \frac{o(h,u)}{\|h\|} = 0$. Величина $dJ = (J'(u), h)$ называется дифференциалом функции $J(\cdot)$ в точке u, а вектор $J'(u) \in \mathcal{R}_n$ – производной, или градиентом этой функции. Если элемент $u \in \mathcal{R}_n$ задан своими координатами $(u_1, u_2, \ldots, u_n)$ в некотором ортонормированном базисе пространства $\mathcal{R}_n$, то координаты вектора производной $J'(u) \in \mathcal{R}_n$ в том же базисе есть частные производные функции $J(u_1, u_2, \ldots, u_n)$, $(u_1, u_2, \ldots, u_n) \in \mathcal{R}_n$, по соответствующим переменным.

Функция $J(\cdot)$ называется дважды дифференцируемой в точке $u \in \mathcal{R}_n$, если, наряду с градиентом $J'(u) \in \mathcal{R}_n$ существует линейный оператор $J''(u)$, действующий из $\mathcal{R}_n$ в $\mathcal{R}_n$ (что обозначается $J''(u) \in (\mathcal{R}_n \to \mathcal{R}_n)$), такой, что для любой точки $h \in \mathcal{R}_n$, для которой $(u+h) \in O_\varepsilon(u)$, функция $J(\cdot)$ представима в виде

$$J(u+h) = J(u) + (J'(u), h) + \frac{1}{2}(J''(u)h, h) + o(\|h\|^2, u).$$

Классический метод основан на утверждении, что в точке u_* минимума дважды дифференцируемой на всем $\mathcal{R}_n$ функции $J(\cdot)$ необходимо, чтобы ее градиент обращался в нуль, а квадратичная форма $(J''(u_*)h, h)$ – неотрицательно определена. Метод состоит в том, чтобы определить все точки, подозрительные на минимум, и выбрать из них те, в которых этот минимум достигается.

Если функция $J(\cdot)$ минимизируется не на всем пространстве, а лишь на некотором его подмножестве $\mathcal{U} \subset \mathcal{R}_n$, то минимум может достигаться и на границе. В этом случае необходимо применение специальных методов минимизации, о которых речь пойдет ниже, в части, посвященной минимизации функций с ограничениями.

Численные методы минимизации.

Основные трудности минимизации функций нескольких переменных можно рассмотреть на примере функций двух переменных $J(u_1, u_2)$, $(u_1, u_2) \in R_2$. Графиком ее является некоторая поверхность в трехмерном пространстве (u_1, u_2, J). Задача на минимум при этом эквивалентна поиску "низшей"точки этой поверхности.

Рельеф этой поверхности можно изобразить линиями уровня, представляющими собой множества точек плоскости, в которых функция $J(\cdot, \cdot)$ принимает заданное значение. По виду линий уровня условно можно выделить три типа рельефов: котловинный, овражный и неупорядоченный.

При котловинном рельефе линии уровня похожи на эллипсы. В малой окрестности точки минимума для гладкой функции $J(\cdot,\cdot)$ рельеф всегда котловинный. Действительно, гладкость в данном случае означает, что в окрестности точки минимума функция $J(\cdot,\cdot)$ дважды дифференцируема, для нее необходимые условия экстремума в точке (x_*, y_*) дают равенства $J'_{u_1}(u_{1*}, u_{2*}) = 0$, $J'_{u_2}(u_{1*}, u_{2*}) = 0$, а разложение по формуле Тейлора дает выражение

$$J(u_1,u_2) = J(u_{1*},u_{2*}) + 1/2J''_{u_1,u_1}(u_{1*},u_{2*})(u_1-u_{1*})^2 +$$
$$+J''_{u_1,u_2}(u_{1*},u_{2*})(u_1-u_{1*})(u_2-u_{2*}) +$$
$$+1/2J''_{u_2,u_2}(u_{1*},u_{2*})(u_2-u_{2*})^2 + o((u_1-u_{1*})^2+(u_2-u_{2*})^2),$$

причем квадратичная форма $1/2J''_{u_1,u_1}(u_{1*},u_{2*})(u_1-u_{1*})^2 + J''_{u_1,u_2}(u_{1*},u_{2*})(u_1-u_{1*})(u_2-u_{2*}) + 1/2J''_{u_2,u_2}(u_{1*},u_{2*})(u_2-u_{2*})^2$ положительно определена. Ее линии уровня выглядят как эллипсы. Случай, когда производные второго порядка обращаются в нуль, по существу ничего нового не дают, так как в этом случае в окрестности минимума линии уровня оказываются кривыми четвертого или более высокого порядка, похожими на эллипсы. Методы численной минимизации, такие, как метод наискорейшего спуска или метод сопряженных направлений, рассчитаны именно на такой тип рельефа, поэтому для их успешного применения начальное приближение важно выбирать как можно ближе к точке (u_{1*}, u_{2*}).

Овражный тип рельефа характеризуется длинными вытянутыми замкнутыми линиями уровня, возможно, с точками излома. Избавится от такого рельефа иногда удается заменой переменных. В противном случае требуется выбирать траекторию, на которой минимизируемая функция убывает, тщательно приспосабливаясь к профилю оврага.

Неупорядоченный тип рельефа характеризуется наличием множества максимумов, минимумов и седловых точек. Рекомендации здесь — такие же, как и для функций с овражным типом рельефа.

Спуск по координатам. Выбирается нулевое приближение $u_1 = u_{1,0}$, $u_2 = u_{2,0}$. Фиксируется значение $u_2 = u_{2,0}$, тогда функция $J(\cdot, u_{2,0})$ зависит только от одного переменного u_1. Найдем ее минимум, используя методы предыдущей главы — пусть он достигается в точке $u_1 = u_{1,1}$. Зафиксируем теперь значение $u_1 = u_{1,1}$ и рассмотрим функцию одного переменного $J(u_{1,1}, \cdot)$. Минимизируем ее как функцию одного переменного и найдем точку минимума $u_2 = u_{2,1}$. На этом завершается первый цикл минимизации; циклы повторяются до тех пор, пока не будут изменяться точки минимума.

Заметим, что этот способ не всегда приводит к успеху — примером может служить функция, линии уровня которой изображены на рис.:

Если же функция достаточно гладкая — например, имеет котлованную структуру, то процесс покоординатного спуска приводит к успеху, хотя скорость сходимости его невелика.

Наискорейший спуск . Этот метод отличается от спуска по координатам тем, что в качестве направления, по которому осуществляется минимизация, выбирается не координатная ось в $\mathcal{R}_n$, а направление, противоположное градиенту функции — то есть направления наибольшей скорости ее убывания. Изменение направления спуска производится либо на каждом шаге, либо при достижении минимума по выбранному направлению. Этот метод сложнее покоординатного спуска, но обладает теми же недостатками.

Метод сопряженных направлений. Как метод покоординатного спуска, так и метод наискорейшего спуска для достижения минимума дважды дифференцируемой функции требуют, вообще говоря, бесконечного числа итераций. Однако можно построить такие направления спуска, что для квадратичной функции

$$J(u) = (u, Au) + (b, u) + c, \quad u,\ b \in \mathcal{R}, \quad A \in (\mathcal{R} \to \mathcal{R}), \qquad (*)$$

процесс спуска сойдется к точному минимуму за конечное число шагов.

Введем норму вектора $u \in \mathcal{R}$, воспользовавшись положительно определенным оператором A:

$$\|u\|_A^2 = (u, Au).$$

Векторы, ортогональные в смысле скалярного произведения $(\cdot, \cdot)_A$, называют сопряженными. Линии уровня квадратичной функции (u, Au) в пространстве с новой метрикой, согласованной со скалярным произведением $(\cdot, \cdot)_A$, очевидно, являются окружностями, а значит, двигаясь по направлению наибольшего спуска в таком пространстве, мы попадем в точку минимума, не меняя направления спуска. Перейти к линиям уровня в виде окружностей с центром в начале координат для функций вида (*) можно с помощью афинной замены переменных $u \to b_0 + Vu$.

Поясним формально, в чем состоит идея методов спуска по сопряженным направлениям. Пусть в пространстве с нормой $\|\cdot\|_A$ найден базис из сопряженных векторов $\{u_i\}$. Выберем произвольную точку u_0 и любое движение из нее представим в базисе, составленном из сопряженных векторов, в виде

$$u = u_0 + \sum_{i=1}^{n} \alpha_i u_i.$$

Подставляя это выражение в выражение для функции (*), найдем

$$J(u) = J(u_0) + \sum_{i=1}^{n} (\alpha_i^2 + 2\alpha_i (u_i, Au_0) + \alpha_i (u_i, b)).$$

Последняя сумма состоит из групп слагаемых, каждая из которых зависит только от одного неизвестного α_i, $i = 1, \ldots, n$, и может минимизироваться по α_i, $i = 1, \ldots, n$, не зависимо от других. Выполняя эту минимизацию, мы найдем точное значение минимума функции (*) за конечное число шагов.

Глава 3. Минимизация функций в гильбертовом пространстве.

Элементы топологии. Обозначим $\mathcal{R}$ — гильбертово пространство, то есть полное нормированное сепарабельное пространство, в котором для любых двух элементов $u, v \in \mathcal{R}$ определено скалярное произведение $(u, v) \in \mathcal{R}_1$, согласованное с нормой: $(u, u) = \|u\|^2$. Напомним, что полным пространством называется такое, в котором любая фундаментальная последовательность $\{u_n\}$ элементов из $\mathcal{R}$ сходится к элементу из $\mathcal{R}$, то есть если $\|u_n - u_p\| \to 0$ при $n, p \to \infty$, то найдется элемент $u_0 \in \mathcal{R}$, такой, что $u_0 = \lim\limits_{n\to\infty}$. Сепарабельность означает, что в $\mathcal{R}$ существует счетное всюду плотное подмножество, и в частности, в таких пространствах можно выбрать полную счетную систему ортонормированных векторов — ортонормированный базис $\mathcal{R}$.

Множество $\mathcal{U} \subset \mathcal{R}$ называется ограниченным, если найдется такое число M, что $\|u\| < M$ при любом $u \in \mathcal{U}$.

Опишем топологию гильбертова пространства, то есть укажем все открытые подмножества $\mathcal{R}$. Для этого определим открытый шар $O_\varepsilon(u_0)$ радиуса $\varepsilon > 0$ с центром в точке u_0 как множество элементов $u \in \mathcal{R}$, для которых норма разности $u - u_0$ не превосходит ε:

$$O_\varepsilon(u_0) = \{u \in \mathcal{R} : \ \|u - u_0\| < \varepsilon\}.$$

Открытый шар в $\mathcal{R}$ радиуса ε назовем ε-окрестностью точки $u_0 \in \mathcal{R}$.

Множество $\mathcal{U} \subset \mathcal{R}$ называется открытым в $\mathcal{R}$, если любой элемент u из $\mathcal{U}$ содержится в $\mathcal{U}$ вместе с некоторой своей ε-окрестностью, то есть для любого $u \in \mathcal{U}$ найдется такое число $\varepsilon > 0$, что $O_\varepsilon(u_0) \subset \mathcal{U}$.

Точка $u \in \mathcal{R}$ называется предельной точкой множества $\mathcal{U}$, если существует последовательность $\{u_n\}$ элементов множества $\mathcal{U}$, сходящаяся к точке u.

Множество $\mathcal{U} \subset \mathcal{R}$ называется замкнутым, если $\mathcal{U}$ содержит все свои предельные точки. Все гильбертово пространство $\mathcal{R}$ и пустое множество $\emptyset$ и открыты, и замкнуты одновременно (по определению).

Множество $\mathcal{U} \subset \mathcal{R}$ называется компактным, если любая последовательность $\{u_n\}$ элементов множества $\mathcal{U}$ содержит подпоследовательность, сходящуюся к точке множества $\mathcal{U}$.

Замыканием $\overline{\mathcal{U}}$ множества $\mathcal{U}$ называется множество элементов гильбертова пространства $\mathcal{R}$, состоящее из элементов $\mathcal{U}$ и всех предельных точек множества $\mathcal{U}$. $\overline{\mathcal{U}}$ — минимальное (по включению) замкнутое множество, содержащее $\mathcal{U}$.

Множество $\mathcal{U} \subset \mathcal{R}$ плотно в $\mathcal{W} \subset \mathcal{R}$, если $\overline{\mathcal{U}} \supseteq \mathcal{W}$.

Точка $u \in \mathcal{U}$ называется изолированной точкой $\mathcal{U}$, если существует ее ε-окрестность $O_\varepsilon(u)$, не содержащая элементов множества $\mathcal{U}$, за исключением u. Замыкание $\overline{\mathcal{U}}$ множества $\mathcal{U}$ состоит из предельных и изолированных точек $\mathcal{U}$.

Множество $\mathcal{U}$ замкнуто тогда и только тогда, когда $\overline{\mathcal{U}} = \mathcal{U}$, или тогда и только тогда, когда открыто множество $\mathcal{R} \setminus \mathcal{U}$.

Пересечение любого числа и объединение конечного числа замкнутых множеств замкнуто.

Множество $\operatorname{int}\mathcal{U}$, состоящее из точек множества $\mathcal{U}$, содержащихся в $\mathcal{U}$ вместе с некоторой своей окрестностью, называется внутренностью $\mathcal{U}$. Множество $\operatorname{int}\mathcal{U}$ является максимальным (по включению) открытым множеством, содержащимся в $\mathcal{U}$. $\mathcal{U}$ открыто тогда и только тогда, когда $\mathcal{U} = \operatorname{int}\mathcal{U}$, и тогда и только тогда, когда $\mathcal{R} \setminus \mathcal{U}$ замкнуто.

Трудности решения задач на минимум функции, заданной на подмножестве $\mathcal{U}$ гильбертова пространства $\mathcal{R}$, связано с тем, что в $\mathcal{R}$ замкнутый шар не является компактным множеством. Действительно, если $\mathcal{U} = \{u : \|u\| \le 1$, и $\{e_k\}$ — ортонормированный базис $\mathcal{R}$, то из последовательности нельзя выбрать сходящейся подпоследовательности, так как $\|e_k - e_n\|^2 = 2$ для любых номеров k и n. Требование компактности множества $\mathcal{R}$ в гильбертовом пространстве является, таким образом, весьма ограничительным. В связи с этим фактом важную роль в гильбертовом пространстве играет так называемая слабая топология.

Определение. Последовательность $\{u_k\}$ элементов гильбертова пространства $\mathcal{R}$ называется слабо сходящейся к элементу $u_0 \in \mathcal{R}$, а u_0 — слабым пределом последовательности $\{u_k\}$, если для любого $v \in \mathcal{R}$ числовая последовательность $\{(u_k, v)\}$ сходится к (u_0, v).

Гильбертово пространство $\mathcal{R}$ слабо полно, то есть условие: для любого $v \in \mathcal{R}$ $\lim\limits_{n,k\to\infty}(u_n - u_k, v) = 0$ необходимо и достаточно для того, чтобы последовательность $\{u_k\} \in \mathcal{R}$ слабо сходилась к некоторому элементу из $\mathcal{R}$. Слабый предел — единственен. Слабо сходящаяся последовательность ограничена по норме $\mathcal{R}$.

Сходящаяся последовательность является и слабо сходящейся. Обратное, вообще говоря, не верно. Действительно, пусть последовательность $\{u_k\}$ слабо сходится к u_0, тогда

$$\|u_k - u_0\|^2 = \|u_k\|^2 - 2(u_k, u_0) + \|u_0\|^2 \to \lim_{k\to\infty}\|u_k\|^2 - \|u_0\|^2,$$

то есть $\{u_k\}$ сходится к u_0 по норме тогда и только тогда, когда числовая последовательность норм $\|u_k\|$ сходится к $\|u_0\|$ при $k \to \infty$.

Приведем пример слабо сходящейся последовательности, не сходящейся по норме. Пусть $\{e_k\} \subset \mathcal{R}$ — ортонормированный базис $\mathcal{R}$, тогда для любого элемента $v \in \mathcal{R}$ скалярное произведение (v, e_k) стремится к нулю при $k \to \infty$ в силу равенства Бесселя

$$\sum_{k=1}^{\infty}(v, e_k)^2 = \|v\|^2 < \infty,$$

тем самым последовательность, составленная из элементов ортонормированного базиса $\mathcal{R}$, слабо сходится к нулевому элементу из $\mathcal{R}$, в то время, как норма каждого элемента последовательности $\{e_k\}$ равна единице.

Определение. Множество $\mathcal{U} \subset \mathcal{R}$ слабо компактно, если из любой последовательности элементов $\mathcal{U}$ можно выбрать подпоследовательность, слабо сходящуюся к элементу из $\mathcal{U}$.

Роль слабой компактности в задачах на минимум функций в гильбертовом пространстве состоит в том, что ограниченное слабо замкнутое множество в $\mathcal{R}$ слабо компактно. В частности, единичный шар $\{u \in \mathcal{R} : \ \|u\| \le 1\}$ слабо компактен в $\mathcal{R}$.

Заметим, что из слабой замкнутости множества $\mathcal{U} \subset \mathcal{R}$ следует его замкнутость. Действительно, слабая замкнутость $\mathcal{U}$ означает, что все его слабо предельные точки содержатся в $\mathcal{U}$. Но предельная точка $\mathcal{U}$ (в смысле сильной сходимости, то есть по норме $\mathcal{R}$ является в то же время и его слабо предельной точкой, и значит, в силу слабой замкнутости, принадлежит $\mathcal{U}$.

Так же, как для конечномерных евклидовых пространств, в гильбертовом пространстве определяются понятия верхнего и нижнего предела, (слабой) полунепрерывности снизу и сверху функции $J(\cdot)$, заданной на $\mathcal{U} \subset \mathcal{R}$. Заметим, что полунепрерывность снизу (сверху) следует из слабой полунепрерывности снизу (сверху) функции $J(\cdot)$.

Будем говорить, что последовательность $\{u_k\}$ слабо сходится к множеству $\mathcal{U} \subset \mathcal{R}$, если все слабо предельные точки этой последовательности принадлежат $\mathcal{U}$.

Условия существования минимума.

Теорема 1. (Вейерштрасса). Пусть функция $J(\cdot)$ определена, конечна и (слабо) полунепрерывна снизу на (слабо) компактном множестве $\mathcal{U} \subset \mathcal{R}$. Тогда $J_* = \inf\limits_{u \in \mathcal{U}} \{J(u)\} > -\infty$, множество точек минимума $\mathcal{U}_* = \{u \in \mathcal{U} : \ J(u) = J_*\}$ не пусто, (слабо) компактно, и любая минимизирующая последовательность (слабо) сходится к $\mathcal{U}_*$.

Доказательство. Пусть $\{u_k\} \subset \mathcal{U}$ — последовательность, минимизирующая функцию $J(\cdot)$ на множестве $\mathcal{U}$, то есть $\lim\limits_{k \to \infty} J(u_k) = J_*$. Так как $\mathcal{U}$ — (слабо) компактно, то существует (слабо) сходящаяся подпоследовательность $\{u_{k_n}\} \subset \{u_k\}$, и $u_* \in \mathcal{U}$ — ее слабый предел. Тогда

$$J_* \le J(u_*) \le \liminf_{k \to \infty} J(u_{n_k}) = \lim_{n \to \infty} J(u_n) = J_*,$$

то есть $J_* > -\infty$, $\mathcal{U}_* \ne \emptyset$ и любая (слабо) предельная точка последовательности $\{u_k\}$ принадлежит множеству $\mathcal{U}_*$, то есть $\{u_k\}$ слабо сходится к $\mathcal{U}_*$.

Если $\mathcal{U}$ — компактно, то сходимость $\{u_k\}$ к $\mathcal{U}_*$ следует из непрерывности расстояния от точки u до множества $\mathcal{U}$ как функции $u \in \mathcal{R}$.

Докажем теперь, что $\mathcal{U}_*$ — (слабый) компакт. Пусть $\{u_n\} \in \mathcal{U}_*$. Так как $\mathcal{U}_* \subset \mathcal{U}$, а $\mathcal{U}$ — (слабо) компактно, то последовательность $\{u_n\}$ имеет (слабо) предельные точки, причем все они принадлежат $\mathcal{U}_*$, так как $\{u_n\}$ — минимизирующая последовательность.

Требование (слабой) компактности множества $\mathcal{U}$ может оказаться слишком сильным, ослабить его можно, наложив ограничение на качество минимизируемой функции $J(\cdot)$.

Выпуклые задачи на минимум. Напомним, что выпуклым называется такое множество $\mathcal{U} \subset \mathcal{R}$, которое вместе с любыми двумя элементами $u, v \in \mathcal{U}$ содержит в себе отрезок $\{x \in \mathcal{R} : \ x = \alpha u + (1-\alpha)v, \ \forall \alpha \in [0,1]\}$. Пустое множество $\emptyset$ и все пространство $\mathcal{R}$ выпуклы по определению.

Перечислим свойства выпуклых множеств.

1. Пересечение любого числа выпуклых множеств — выпукло.

2. Пусть $\mathcal{U}_i \subset \mathcal{R}$, $i = 1, \ldots, m$, и $\mathcal{U} = \mathcal{U}_1 + \mathcal{U}_2 + \ldots + \mathcal{U}_m \equiv \sum\limits_{i=1}^{m} \mathcal{U}_i$ — векторная сумма множеств, то есть множество элементов вида $u = u_1 + u_2 + \ldots + u_m$, $u_i \in \mathcal{U}_i$, $i = 1, \ldots, m$, а $\lambda_i \mathcal{U}_i$ — множество элементов вида $u = \lambda_i u_i$, $\lambda \in \mathcal{R}_1$, $u_i \in \mathcal{U}_i$, $i = 1, \ldots, m$. Тогда, если $\mathcal{U}_i$, $i = 1, \ldots, m$, выпуклы, то для любого набора чисел λ_i, $i = 1, \ldots, m$, множество $\sum\limits_{i=1}^{m} \lambda_i \mathcal{U}_i$ является выпуклым.

Действительно, пусть $u, v \in \mathcal{U}$, тогда найдутся такие элементы $u_i \in \mathcal{U}_i$, $v_i \in \mathcal{U}_i$, что $u = \sum\limits_{i=1}^{m} \lambda_i u_i$, $v = \sum\limits_{i=1}^{m} \lambda_i v_i$, и

$$\alpha u + (1-\alpha)v = \sum_{i=1}^{m} (\lambda_i \alpha u_i + (1-\alpha)\lambda_i v_i) = \sum_{i=1}^{m} \lambda_i (\alpha u_i + (1-\alpha)v_i) \in \mathcal{U},$$

так как в силу выпуклости $\mathcal{U}_i$ $\alpha u_i + (1-\alpha)v_i \in \mathcal{U}_i$, $i = 1, \ldots, m$.

Как следствие этого, для выпуклых множеств $\mathcal{U}_1$ и $\mathcal{U}_2$ их векторная разность $\mathcal{U}_1 - \mathcal{U}_2 = \mathcal{U}_1 + (-1)\mathcal{U}_2$ тоже выпукла.

3. Если выпуклое множество содержит более одной точки, оно не может содержать изолированных точек.

4. Если $\mathcal{U}$ выпукло, то $\overline{\mathcal{U}}$ и $\operatorname{int}\mathcal{U}$ также выпуклы.

5. Пересечение всех выпуклых множеств, содержащих $\mathcal{U}$, называемое выпуклой оболочкой $\mathcal{U}$ и обозначаемое $\operatorname{co}\mathcal{U}$, является минимальным по включению выпуклым множеством, содержащим $\mathcal{U}$. Если $\mathcal{U}$ — замкнуто и ограничено, то $\operatorname{co}\mathcal{U}$ замкнуто, ограничено и выпукло.

Заметим, что если $\mathcal{U}$ целиком лежит в конечномерном подпространстве $\mathcal{R}$, то

$$\operatorname{co}\mathcal{U} = \left\{ u \in \mathcal{R} : u = \sum_{k=1}^{M} \alpha_k u_k, \ \alpha_k \geq 0, \ u_k \in \mathcal{U}, k = 1, \ldots, M, \ \sum_{k=1}^{M} \alpha_k = 1, \right\},$$

где M — любое целое число. Иными словами, любая точка множества $\operatorname{co}\mathcal{U}$ представима в виде выпуклой комбинации *не более, чем конечного числа точек* из множества $\mathcal{U}$.

6. Теорема Мазура. Выпуклое замкнутое множество слабо замкнуто. Следствием этой теоремы является то, что ограниченное выпуклое замкнутое множество слабо компактно.

Определение. Функция $J(\cdot)$ выпукла на выпуклом множестве $\mathcal{U} \subset \mathcal{R}$, если для любых $u, v \in \mathcal{U}$ и для любого числа $\alpha \in [0,1]$

$$J(\alpha u + (1-\alpha)v) \leq \alpha J(u) + (1-\alpha)J(v).$$

Если равенство возможно только при $\alpha = 0$ или $\alpha = 1$, то $J(\cdot)$ строго выпукла. Если $\mathcal{U} = \emptyset$, то любая функция на $\mathcal{U}$ выпукла по определению.

Перечислим свойства выпуклых функций.

1. Если функция $J(\cdot)$ выпукла на $\mathcal{U}$, а $F(\cdot)$ выпукла и монотонно не убывает на $\mathcal{R}_1$, то их композиция $F(J(\cdot))$ выпукла на $\mathcal{U}$.

Действительно,

$$F(J(\alpha u + (1-\alpha)v)) \le F(\alpha J(u) + (1-\alpha)J(v)) \le \alpha F(J(u)) + (1-\alpha)F(J(v))$$

для всех $u, v \in \mathcal{U}$ и $0 \le \alpha \le 1$.

2. Для любого семейства $J_i(\cdot)$, $i \in I$, выпуклых на $\mathcal{U}$ функций функция

$$J(u) = \sup\{J_i \mid i \in I\}$$

выпукла на $\mathcal{U}$.

3. Конечная линейная комбинация $\alpha_1 J_1(\cdot) + \alpha_2 J_2(\cdot) + \ldots + \alpha_m J_m(\cdot)$ выпуклых на $\mathcal{U}$ функций $J_i(\cdot)$, $i = 1, \ldots, m$, также выпукла.

Лемма 1. Пусть $J(\cdot)$ определена и выпукла на всем пространстве $\mathcal{R}$ и для любого числа α и любого элемента $u \in \mathcal{R}$ $J(\alpha u) = |\alpha| J(u)$. Тогда для любых элементов $u, v \in \mathcal{R}$

1. $J(u+v) \le J(u) + J(v)$;
2. $J(u) = J(-u)$;
3. $J(u) = \frac{1}{2}(J(u) + J(-u)) \ge 0$.

Доказательство. Докажем утверждение пункта 3. Для любых $u, v \in \mathcal{R}$

$$\frac{1}{2}J(u+v) = J\left(\frac{1}{2}u + \frac{1}{2}v\right) \le J\left(\frac{1}{2}u\right) + J\left(\frac{1}{2}v\right) = \frac{1}{2}(J(u) + J(v)).$$

Лемма 2. Пусть $J(\cdot)$ определена, выпукла и полунепрерывна снизу на всем пространстве $\mathcal{R}$ и для любого числа α и любого элемента $u \in \mathcal{R}$ выполнено равенство $J(\alpha u) = |\alpha| J(u)$. Тогда $J(\cdot)$ непрерывна на $\mathcal{R}$ и существует такое число $m > 0$, что

$$J(u) \le m\|u\|$$

для любого элемента $u \in \mathcal{R}$.

Доказательство. Покажем, что существует такое число $m > 0$, что для любого u из единичного шара $O_1(0) = \{v \in \mathcal{R} : \ \|v\| < 1\}$ выполняется неравенство $J(u) \le m$. Предположим, что это не так, то есть $J(\cdot)$ неограничена в $O_1(0)$. Тогда в силу свойств $J(\cdot)$ она неограничена и в любом другом шаре $O_\varepsilon(u_0)$, так как если $J(u) \le n$ для $u \in O_\varepsilon(u_0)$, то

$$J\left(\frac{u-u_0}{\varepsilon}\right) \le \frac{1}{\varepsilon}\left(J(u) + J(u_0)\right) \le \frac{2n}{\varepsilon}$$

для $\frac{u-u_0}{\varepsilon} \in O_1(0)$. В силу предположения найдется элемент $u_1 \in O_1(0)$, такой, что $J(u_1) > 1$. Так как $J(\cdot)$ — полунепрерывна снизу, то найдется и окрестность $O_{\varepsilon_1}(u_1) \subset O_1(0)$, такая, что неравенство $J(u) > 1$ выполняется для всех

$u \in O_{\varepsilon_1}(u_1)$, и $\varepsilon_1 < \frac{1}{2}$. Поскольку $J(\cdot)$ неограничена и в этом шаре, то найдется элемент $u_2\ in O_{\varepsilon_1}(u_1)$ и, соответственно, шар $O_{\varepsilon_2}(u_2)$, во всех точках которого $J(u) > 2$, и $\varepsilon_2 < \frac{1}{2^2}$. Продолжая процедуру выбора шаров, получим последовательность вложенных шаров $O_{\varepsilon_1}(u_1) \supset O_{\varepsilon_2}(u_2) \supset O_{\varepsilon_3}(u_3) \supset \ldots$, таких, что $J(u) > n$, как только $u \in O_{\varepsilon_n}(u_n)$, а $\varepsilon_n < \frac{1}{2^n}$. Последовательность центров шаров сходится к точке $u_* \in \mathcal{R}$, и $J(u_k) \to \infty$ при $k \to \infty$, что невозможно в силу предположений о свойствах функции $J(\cdot)$. Значит, найдется такое число $m > 0$, что $J(u) \le m\|u\|$ для всех $u \in \mathcal{R}$.

Покажем, что $J(\cdot)$ непрерывна. Пусть $\{u_k\} \to u_0$ при $k \to \infty$, тогда $J(u_0) \le \liminf\limits_{k\to\infty} J(u_k)$, и

$$J(u_k) = J(u_0 + (u_k - u_0)) \le J(u_0) + J(u_k - u_0) \le J(u_0) + m\|u_k - u_0\| \to_{k\to\infty} J(u_0),$$

поэтому $\limsup\limits_{k\to\infty} J(u_k) \le J(u_0)$, то есть $J(\cdot)$ полунепрерывна и сверху, а следовательно, и непрерывна.

Теорема 2. Пусть $J_i(\cdot)$, $i \in I$, — множество определенных на $\mathcal{R}$ выпуклых функций, удовлетворяющих условию $J_i(\alpha u) = |\alpha| J_i(u)$, для всех $u \in \mathcal{R}$, $i \in I$, $\alpha \in \mathcal{R}_1$. Если $J_i(\cdot)$, $i \in I$ полунепрерывны снизу и числовое множество $\{J_i(u)\}$ ограничено при любом $u \in \mathcal{R}$, то $J(u) = \sup\limits_{i\in I}\{J_i(u)\}$, $u \in \mathcal{R}$, выпукла и непрерывна на $\mathcal{R}$, причем для некоторого $m > 0$ неравенство $J(u) \le m\|u\|$ выполнено для всех $u \in \mathcal{R}$.

Доказательство. Очевидно, $J(\cdot)$ выпукла и $J(\alpha u) = |\alpha| J(u)$, для всех $u \in \mathcal{R}$, $\alpha \in \mathcal{R}_1$. Покажем, что $J(\cdot)$ полунепрерывна снизу на $\mathcal{R}$. В силу определения $J(\cdot)$, для любого фиксированного $u \in \mathcal{R}$ и для любого $\varepsilon > 0$ найдется такое $i \in I$, что

$$J(u_0) - J_i(u_0) < \varepsilon.$$

В силу полунепрерывности снизу функций $J_i(\cdot)$ найдется такое $\delta > 0$, такое, что для всех u, таких, что $\|u - u_0\| \le \delta$, выполнено неравенство $|J_i(u_0) - J_i(u)| < \varepsilon$. Следовательно,

$$J(u) - J(u_0) > J(u) - J_i(u_0) - \varepsilon \ge J_i(u) - J_i(u_0) - \varepsilon > -2\varepsilon,$$

то есть $J(\cdot)$ — полунепрерывна снизу, а значит, в силу Леммы 2, непрерывна и неравенство

$$J(u) \le m\|u\|$$

выполнено для некоторого $m > 0$ и для всех $u \in \mathcal{R}$.

В экстремальных задачах выпуклость функции $J(\cdot)$ позволяет сформулировать достаточные условия минимума.

Теорема 3. Пусть функция $J(\cdot)$ выпукла на выпуклом множестве $\mathcal{U} \subset \mathcal{R}$. Тогда всякая точка локального минимума функции $J(\cdot)$ на $\mathcal{U}$ является и точкой глобального минимума, причем множество $\mathcal{U}_* = \{u \in \mathcal{U} : \ J(u) = J_*\}$ выпукло. Если $J(\cdot)$ строго выпукла на $\mathcal{U}$, то $\mathcal{U}_*$ содержит не более одной точки.

Напомним, что $u_0 \in \mathcal{U}$ является точкой локального минимума функции $J(\cdot)$, если существует ε-окрестность $O_\varepsilon(u_0)$ точки u_0, такая, что $J(u_0) \le J(u)$ для всех $u \in O_\varepsilon(u_0) \cap \mathcal{U}$.

Доказательство. Пусть u_* — точка локального минимума, и u — произвольная точка из $\mathcal{U}$. Тогда в силу выпуклости $\mathcal{U}$ найдется достаточно малое число $\alpha > 0$, такое, что $u_* + \alpha(u - u_*) \in O_\varepsilon(u_0) \cap \mathcal{U}$, и поэтому

$$J(u_*) \le J(u_* + \alpha(u - u_*)) \le (1-\alpha)J(u_*) + \alpha J(u) = J(u_*) + \alpha(J(u) - J(u_*)),$$

где последнее неравенство выполнено в силу выпуклости $J(\cdot)$. Сравнивая левые и правые части последнего соотношения, получим, что $J(u) \le J(u_*)$, то есть u_* — точка глобального минимума.

Если $u_1, u_2 \in \mathcal{U}_*$, то для любого $\alpha \in [0,1]$ $J_* \le J(\alpha u_1 + (1-\alpha)u_2) \le \alpha J(u_1) + (1-\alpha)J(u_2) = J_*$, то есть для любого $\alpha \in [0,1]$ $\alpha u_1 + (1-\alpha)u_2 \in \mathcal{U}_*$, что означает выпуклость множества $\mathcal{U}_*$.

Если $J(\cdot)$ строго выпукла, то для $u_1 \ne u_2$, $u_1, u_2 \in \mathcal{U}_*$, и для любого $\alpha \in [0,1]$ $J_* \le J(\alpha u_1 + (1-\alpha)u_2) < \alpha J(u_1) + (1-\alpha)J(u_2) = J_*$, чего не может быть, следовательно, $u_1 = u_2$.

Однако в такой общей задаче $\mathcal{U}_*$ может оказаться пустым, причем даже если множество $\mathcal{U}$ замкнуто. В качестве примера приведем функцию $J(u) = \frac{1}{\|u\|+1}$, определенную на одномерном выпуклом замкнутом подпространстве $\mathcal{U}_e = \{u = \beta e, -\infty < \beta < \infty\} \subset \mathcal{R}$, $e \in \mathcal{R}$ — фиксированный вектор.

Для того, чтобы функция $J(\cdot)$ достигала на $\mathcal{U}$ своей точной нижней грани, нужны более сильные требования на функцию $J(\cdot)$ и на множество $\mathcal{U}$. В частности, если $\mathcal{U}$ — выпукло, замкнуто и ограничено, то, в силу теоремы Мазура, оно слабо компактно, а если $J(\cdot)$ — выпукла и полунепрерывна снизу, то она и слабо полунепрерывна снизу. Тогда для таких $J(\cdot)$ и $\mathcal{U}$ выполнены условия теоремы Вейерштрасса, и значит, $J_* = \inf\limits_{u \in \mathcal{U}} \{J(u)\} > -\infty$, множество точек минимума $\mathcal{U}_* = \{u \in \mathcal{U} : \ J(u) = J_*\}$ не пусто, слабо компактно, и любая минимизирующая последовательность слабо сходится к $\mathcal{U}_*$.

Так же, как и для конечномерного случая, можно отказаться от ограниченности множества $\mathcal{U}$, наложив ограничение на $J(\cdot)$.

Теорема 4. Пусть функция $J(\cdot)$ выпукла и полунепрерывна снизу на выпуклом замкнутом множестве $\mathcal{U} \subset \mathcal{R}$, и для любого $u_0 \in \mathcal{U}$ лебегово множество

$$\Lambda_{J(u_0)} \equiv \{u \in \mathcal{U} : J(u) \le J(u_0)\}$$

ограничено. Тогда $J_* = \inf\limits_{u \in \mathcal{U}} \{J(u)\} > -\infty$, множество точек минимума $\mathcal{U}_* = \{u \in \mathcal{U} : \ J(u) = J_*\}$ не пусто, выпукло, замкнуто и ограничено, и любая минимизирующая последовательность слабо сходится к $\mathcal{U}_*$.

Укажем класс функций, для которых выполняется сходимость по норме минимизирующей последовательности $\{u_k\}$ к множеству $\mathcal{U}_*$ при $k \to \infty$.

Определение . Функция $J(\cdot)$ называется сильно выпуклой на выпуклом множестве $\mathcal{U} \subset \mathcal{R}$, если для любых $u, v \in \mathcal{U}$, для любого $\alpha \in [0,1]$ и для некоторого числа $q > 0$ выполняется неравенство

$$J(\alpha u + (1-\alpha)v) \le \alpha J(u) + (1-\alpha)J(v) - \alpha(1-\alpha)q\|u-v\|^2.$$

Теорема 5. Пусть функция $J(\cdot)$ сильно выпукла и полунепрерывна снизу на выпуклом замкнутом множестве $\mathcal{U} \subset \mathcal{R}$. Тогда для любого $v \in \mathcal{U}$ лебегово множество

$$\Lambda_{J(v)} \equiv \{u \in \mathcal{U} : J(u) \le J(v)\}$$

выпукло, замкнуто и ограничено, $J_* = \inf\limits_{u\in\mathcal{U}}\{J(u)\} > -\infty$, множество точек минимума $\mathcal{U}_* = \{u \in \mathcal{U} : \ J(u) = J_*\}$ не пусто и состоит из единственной точки $\{u_*\}$, причем любая минимизирующая последовательность $\{u_k\}$ сходится к u_* по норме $\mathcal{R}$ так, что выполнено неравенство

$$q\|u_k - u_*\|^2 \le J(u_k) - J(u_*), \quad k = 1, 2 \ldots .$$

Доказательство. Если U — ограниченное замкнутое множество, то оно слабо компактно (в силу выпуклости и теоремы Мазура), и справедлива теорема Вейерштрасса.

Рассмотрим случай, когда множество U неограничено. Выберем некоторую точку $v \in U$ и рассмотрим шар $S_1(v$ единичного радиуса с центром в точке v. Тогда в силу слабой компактности шара $S_1(v)$ выполнено

$$\inf_{S_1(v)} J(u) = J_S^* > -\infty, \Rightarrow J(u) \ge J_S^* \quad \text{для любого } u \in S_1(v).$$

Обозначим $\nu_v = J(v) - J_S^*$, тогда $J(u) \ge J_S^* = J(v) - \nu_v$.

Выберем теперь произвольную точку U, лежащую вне шара $S_1(v)$: $u \in U \backslash S_1(v)$. Тогда $\|u - v\| > 1$, и

$$0 \le \alpha_0 = \frac{1}{\|u-v\|} < 1,$$

и для любого $\alpha = \alpha_0$ из определения сильно выпуклой функции следует

$$\alpha_0 J(u) \ge J(v + \alpha_0(u-v)) - (1-\alpha_0)J(v) + \alpha_0(1-\alpha_0)q\|u-v\|^2. \qquad (1)$$

Но $\dfrac{\alpha_0}{\|u-v\|} = 1$, поэтому точка $v + \alpha_0(u-v)$ принадлежит сфере — границе шара $S_1(v)$. Следовательно, $v + \alpha_0(u-v) \in S_1(v)$ и $J(v + \alpha_0(u-v)) \ge J(v) - \nu_v$. Продолжим неравенство (1):

$$\alpha_0 J(u) \ge \alpha_0 J(v) - \nu_v + \alpha_0(1-\alpha_0)q\|u-v\|^2.$$

Поделив обе части полученного неравенства на α_0, имеем

$$J(u) \geq J(v) + (1-\alpha_0)q\|u-v\|^2 - \frac{\nu_v}{\alpha_0} =$$

$$= J(v) + q\|u-v\|^2 - (\|u-v\|\sqrt{q})\left(\sqrt{q} + \frac{\nu_v}{\sqrt{q}}\right).$$

Для последнего слагаемого, представляющего собой произведение двух сомножителей, воспользуемся неравенством $ab \leq (a^2+b^2)/2$ и получим

$$J(u) \geq J(v) + q\|u-v\|^2/2 - \frac{(\sqrt{q} + \nu_v/\sqrt{q})}{2}. \tag{2}$$

Напомним, что полученное неравенство доказано для всех $u \in U \backslash S_1(v)$. Но такое же неравенство выполнено и для любого $u \in S_1(v)$. Действительно, если $\|u-v\| \leq 1$, то справедливо неравенство

$$\nu_v < \frac{(\sqrt{q} + \nu_v/\sqrt{q})^2}{2} - \frac{q\|u-v\|^2}{2} = \frac{q}{2} + \frac{\nu_v^2}{2q} + \nu_v - \frac{q\|u-v\|^2}{2}.$$

Отсюда и из определения ν_v следует справедливость неравенства (2) и для всех $u \in S_1(v)$, а в частности, и для любого $u \in \Lambda_{J(v)}$. Из (2) следует, что

$$q(\|u-v\|^2/2) - (\sqrt{q} + \nu_v/\sqrt{q})^2 \leq J(u) - J(v) \leq 0.$$

Поэтому величина $\|u-v\|^2$ ограничена для всех $u \in \Lambda_{J(v)}$, т.е. множество Лебега ограничено.

Выпуклость и замкнутость множества Лебега $\Lambda_{J(v)}$ следует из свойств сильно выпуклых функций.

Заметим, что требования строгой выпуклости $J(\cdot)$ на $\mathcal{U}$ недостаточно для того, чтобы $\mathcal{U}_* \neq \emptyset$. В качестве примера рассмотрим функцию строго выпуклую функцию $J(u) = e^{-u}$, заданную на выпуклом замкнутом множестве $\mathcal{R}_1$: она не достигает своей точной нижней грани, равной нулю, ни в одной точке числовой оси $\mathcal{R}_1$.

Приведем пример сильно выпуклой функции.

Пример 1. Пусть $J(u) = \|Au - z\|^2$, где $u \in \mathcal{R}$, $z \in \mathcal{R}$ — фиксированный элемент, $A \in \mathcal{R} \to \mathcal{R}$ – ограниченный линейный оператор, действующий из $\mathcal{R}$ в $\mathcal{R}$, причем $\|Au\|^2 > \omega\|u\|^2$ для некоторого $\omega > 0$ и всех $u \in \mathcal{U}$. Действительно,

$$\|A(\alpha u + (1-\alpha)v) - z\|^2 = \alpha^2\|Au-z\|^2 + (1-\alpha)^2\|Av-z\|^2 + 2\alpha(1-\alpha)(Au-z, Av-z) =$$

$$= \alpha\|Au-z\|^2 + (1-\alpha)\|Av-z\|^2 + (\alpha^2-\alpha)\|Au-z\|^2 + ((1-\alpha)^2 -$$

$$-(1-\alpha))\|Av-z\|^2 + 2\alpha(1-\alpha)(Au-z, Av-z) =$$

$$= \alpha\|Au-z\|^2 + (1-\alpha)\|Av-z\|^2 - \alpha(1-\alpha)\|Au-Av\|^2 \leq$$

$$\leq \alpha\|Au-z\|^2 + (1-\alpha)\|Av-z\|^2 - \omega\alpha(1-\alpha)\|u-v\|^2.$$

В частности, при $A = I$, $z = 0$ получаем, что квадрат нормы $u \in \mathcal{R}$ — тоже сильно выпуклая функция на $\mathcal{R}$. Воспользовавшись теоремой 5, заметим, что если $\mathcal{U}$ — выпуклое и замкнутое множество, то существует единственный элемент $\mathcal{U}$ с минимальной нормой. Кроме того, так как квадрат нормы — непрерывная функция, то для любого элемента $v \in \mathcal{R}$ существует единственный элемент $u \in \mathcal{U}$, такой, что

$$\|v - u\| = \inf_{z \in \mathcal{U}} \{\|v - z\|\},$$

называемый проекцией элемента $v \in \mathcal{R}$ на $\mathcal{U}$. Если $\mathcal{U} \subset \mathcal{R}$ — линейное подпространство $\mathcal{R}$, то легко показать, что $v - u$ ортогонально любому элементу $z \in \mathcal{U}$, поэтому $u \in \mathcal{U}$ называется ортогональной проекцией элемента $v \in \mathcal{R}$ на $\mathcal{U}$. Отсюда же следует единственность ортогонального разложения гильбертова пространства: любой элемент $v \in \mathcal{R}$ можно единственным образом представить в виде суммы $v = v_1 + v_2$, где $v_1 = u \in \mathcal{U}$, а v_2 принадлежит ортогональному дополнению к $\mathcal{U}$, то есть ортогонален любому вектору из $\mathcal{U}$.

Градиент функции $J(\cdot)$, заданной на гильбертовом пространстве $\mathcal{R}$, и ее вторая производная определяются так же, как для функций, заданных на $\mathcal{R}_n$, $n < \infty$: функция $J(\cdot)$ дифференцируема в точке $u \in \mathcal{R}$, если она определена в ε-окрестности точки $u \in \mathcal{R}$ и для любого $h \in O_\varepsilon(u)$ представима в виде

$$J(u + h) = J(u) + (J'(u), h) + o(u, \|h\|),$$

где $o(u, \|h\|)$ — бесконечно малая функция более высокого порядка малости, чем $\|h\|$ при $h \to 0$. Элемент $J'(u) \in \mathcal{R}$ называется градиентом функции $J(\cdot)$ в точке $u \in \mathcal{U}$, а $(J'(u), h)$ — ее первым дифференциалом в точке $u \in \mathcal{U}$.

Функция $J(\cdot)$ дважды дифференцируема в точке $u \in \mathcal{R}$, если она дифференцируема в u и для любого $h \in O_\varepsilon(u)$ представима в виде

$$J(u + h) = J(u) + (J'(u), h) + \frac{1}{2}(J''(u)h, h) + o(u, \|h\|^2),$$

где $o(u, \|h\|^2)$ — бесконечно малая функция более высокого порядка малости, чем $\|h\|^2$ при $h \to 0$. Линейный оператор $J''(u) \in \mathcal{R} \to \mathcal{R}$ называется второй производной функции $J(\cdot)$ в точке $u \in \mathcal{U}$, а квадратичная форма $(J''(u)h, h)$ — ее вторым дифференциалом в точке $u \in \mathcal{U}$.

Теорема 6. Пусть $\mathcal{U} \subset \mathcal{R}$ — выпуклое множество. Дифференцируемая функция $J(\cdot)$ выпукла на $\mathcal{U}$ тогда и только тогда, когда для любых $u, v \in \mathcal{U}$ выполняется любое из неравенств

$$J(u - J(v) \geq (J'(v), u - v),$$

$$(J'(u) - J'(v)), u - v) \geq 0.$$

Если $J(\cdot)$ дважды дифференцируема на $\mathcal{U}$, и $\operatorname{int}\mathcal{U} \neq \emptyset$, то для выпуклости функции $J(\cdot)$ на $\mathcal{U}$ необходимо и достаточно, чтобы для любого $v \in \mathcal{R}$ и для любого

$u \in \mathcal{U}$ выполнялось неравенство

$$(J''(u)v, v) \geq 0.$$

Необходимые и достаточные условия минимума дифференцируемой функции даются в следующей теореме.

Теорема 7. Пусть функция $J(\cdot)$ задана на множестве $\mathcal{U} \subset \mathcal{R}$, u_* — точка локального минимума функции $J(\cdot)$. Если $J(\cdot)$ дифференцируема в точке u_*, то $J'(u_*) = 0$. Если $J(\cdot)$ дважды дифференцируема в точке u_*, то $J'(u_*) = 0$, $(J''(u_*)v, v) \geq 0$ для любого $v \in \mathcal{R}$, и условие $J''(u_*)v, v) \geq \delta\|v\|^2$, выполняемое для некоторого $\delta > 0$ и всех $v \in \mathcal{R}$, достаточно для того, чтобы $J(\cdot)$ достигала в точке u_* локального минимума.

Приведем пример, свидетельствующий о том, что в условиях теоремы 7 нельзя считать $\delta = 0$.

Пример 2. Пусть $J(u) = \sum\limits_{k=1}^{\infty} \left(\frac{u_k^2}{k^3} - u_k^4\right)$, $u = (u_1, u_2, \ldots) \in l_2$. В точке $u = 0$

$$J'(0) = \left(\frac{2u_1}{1^3} - 4u_1^3, \frac{2u_2}{2^3} - 4u_2^3, \ldots\right)|_{u=0} = 0,$$

и

$$(J''(0)v, v) = \sum_{k=1}^{\infty} 2\frac{v_k^2}{k^3} > 0$$

при $v \neq 0$. Однако точка $u = 0$ не является точкой минимума, так как в сколь угодно малой окрестности нуля найдется элемент $h \in l_2$, такой, что $J(h) < 0$. Например, при $h_n = (\underbrace{0, 0, \ldots, 0}_{n-1}, \frac{1}{n}, 0, \ldots)$, то $J(h_n) - J(0) = \frac{1}{n^5} - \frac{1}{n^4} < 0$ для всех $n > 1$.

Следующая теорема дает достаточные условия минимума функции на выпуклом множестве.

Теорема 8. Пусть $\mathcal{U} \subset \mathcal{R}$ — выпуклое множество, $J(\cdot)$ — дифференцируемая функция $J(\cdot)$ на $\mathcal{U}$. Тогда для любых $u_* \in \mathcal{U}_*$ и $u \in \mathcal{U}$ выполнено неравенство

$$(J'(u_*), u - u_*) \geq 0 \tag{1}$$

при этом для внутренних точек $u_* \in \operatorname{int}\mathcal{U}$ это равенство эквивалентно равенству

$$J'(u_*) = 0.$$

Если, кроме того, $J(\cdot)$ выпукла на $\mathcal{U}$, то выполнения соотношения (1) и достаточно для того, чтобы $u_* \in \mathcal{U}_*$.

В заключение части 1 рассмотрим задачу минимизации функции

$$J(u) = \|Au - z\|^2$$

на гильбертовом пространстве $\mathcal{R}$, где $z \in \mathcal{R}$ — фиксированный элемент, а $A \in \mathcal{R} \to \mathcal{R}$ — заданный вполне непрерывный оператор. Если $z \in \mathcal{R}(A)$, где $\mathcal{R}(A)$ — пространство значений A, то, очевидно, $u = A^{-1}z \in \mathcal{R}$. Однако если $z \notin \mathcal{R}(A)$, то минимизирующая последовательность для $J(\cdot)$ неограничена по норме. Действительно, пусть $\mathcal{R} = l_2$, $e_n = (\underbrace{0, 0, \ldots, 0}_{n-1}, 1, 0, \ldots)$ и $Ae_n = \frac{e_n}{n}$. Если $z = (1, 1/2, 1/3, \ldots, 1/n, \ldots) \in l_2$, то $z \notin \mathcal{R}(A)$, так как формальное решение уравнения $Au = z$, выполненное в базисе собственных векторов оператора A приводит к результату $u_* = (1, 1, \ldots, 1, \ldots)$, но такой элемент не принадлежит l_2. Для того, чтобы сделать задачу поиска точки минимума функции $J(\cdot)$ разрешимой для любого $z \in l_2$, расширяют область определения $J(\cdot)$, пополняя пространство $\mathcal{R}$ по норме $\|\cdot\|_- = \|A\cdot\|$. Пусть $\mathcal{R}_-$ — пополнение $\mathcal{R}$ по норме $\|\cdot\|$. Продолжим функцию $J(\cdot)$ на все $\mathcal{R}_-$ и сохраним за ней прежнее обозначение. Теперь $J(\cdot)$ — сильно выпукла на $\mathcal{R}_-$, и для нее справедлива теорема 5.

Часть 2. Минимизация с ограничениями.

Глава 4. Математическое программирование.

Пусть в $\mathcal{R}_n$ задана функция $J(\cdot)$, и ее минимум ищется на множестве $\mathcal{U} \subseteq \mathcal{R}_n$, в точках которого выполняются условия двух типов:

$$\mathcal{U} = \{u \in \mathcal{R}_n : \ \varphi_i(u) = 0,\ i = 1, \dots, m,\ \ \psi_j(u) \geq 0,\ j = 1, \dots, p\} \qquad (1)$$

Если функции $\varphi_i(u) = 0,\ i = 1, \dots, m$, независимы, то условия типа равенств выделяет в пространстве $\mathcal{R}_n$ $n-m$-мерную поверхность, поэтому число m таких условий должно быть меньше, чем n. Условия типа неравенств выделяют в $\mathcal{R}_n$ n-мерную область, ограниченную гиперповерхностями $\psi_j(u) = 0,\ j = 1, \dots, p$; число p таких условий может быть произвольным.

Задачи поиска минимума функции $J(\cdot)$ на множестве (1) носят название задач математического программирования. В зависимости от вида функций $J(\cdot)$, $\varphi_i(\cdot)$, $i = 1, \dots, m$, $\psi_j(\cdot)$, $j = 1, \dots, p$, различают задачи линейного, выпуклого, нелинейного программирования.

Задачи линейного программирования. В задачах линейного программирования требуется минимизировать линейную функцию на множестве, заданном как решение системы линейных уравнений или неравенств.

Рассмотрим задачу линейного программирования для функций, заданных на n-мерных линейных евклидовых пространствах, $n < \infty$. Пусть $J(u) = (c, u)$, $u \in \mathcal{R}_n$, c — заданный элемент $\mathcal{R}_n$, $n < \infty$, $(\cdot, \cdot)$ — скалярное произведение в $\mathcal{R}_n$, и функция $J(\cdot)$ минимизируется на множестве

$$\mathcal{U} = \{u \in \mathcal{R}_n : (a_i, u) \leq b_i,\ i = 1, \dots, m;\ \ (a_i, u) = b_i,\ i = m+1, \dots, s\}, \qquad (*)$$

где $a_i \in \mathcal{R}_n$, $i = 1, \dots, s$ — заданные элементы, b_i, $i = 1, \dots, s$ — заданные числа, а $0 \leq m \leq s$, то есть не исключается случай, когда условия типа равенств или (и) типа неравенств отсутствуют.

Иногда рассматривают задачу линейного программирования в несколько более подробной постановке. Считают, что задан ортонормированный базис $\{e_k\} \subset \mathcal{R}_n$, и векторы u, c и a_i, $i = 1, \dots, n$, заданы своими координатами в этом базисе. Тогда из условий вида (*) явно выделяют требование положительности координат вектора u; множество $\mathcal{U}$ в этом случае принимает вид

$$\mathcal{U} = \{u \in \mathcal{R}_n,\ u_k \geq 0,\ \ k \in I;\ \ (a_i, u) \leq b_i;\ \ i = 1, \dots, m;$$
$$(a_i, u) = b_i;\ \ i = m+1, \dots, s\}, \qquad (+)$$

где I — некоторое подмножество индексов из множества $\{1, 2, \dots, n\}$. Эквивалентность условий (*) и (+) следует из эквивалентности неравенств

$$u_k \geq 0 \quad \Leftrightarrow \quad (-e_k, u) \leq 0.$$

При заданном ортонормированном базисе часто используют обозначение $x < y$, означающее, что все координаты векторов x и y связаны неравенствами $x_i < y_i$, $i = 1, \ldots, n$. Теперь общую задачу линейного программирования можно записать в виде

$$J(u) = (c, u) \to \inf_{u \in \mathcal{U}},$$

$$\mathcal{U} = \{u \in \mathcal{R}_n,\ u_k \geq 0;\ k \in I;\ \ Au \leq b;\ \overline{A}u = \overline{b}\},$$

где A — матрица размера $m \times n$, k-тая строка которой является вектором a_k, $k = 1, \ldots, m$, а $\overline{A}$ — матрица размера $(s - m) \times n$, j-тая строка которой является вектором a_j, $j = m + 1, \ldots, s$; векторы $b \in \mathcal{R}_m$ и $\overline{b} \in \mathcal{R}_{s-m}$ имеют координаты $b = (b_1, \ldots, b_m)$, $b = (b_{m+1}, \ldots, b_s)$ соответственно.

Точку $u_* \in \mathcal{U}$ назовем точкой минимума функции $J(\cdot) = (c, \cdot)$ на множестве $\mathcal{U}$, если $(c, u_*) = \inf\limits_{u \in \mathcal{U}} J(u) = J_*$.

Из общей задачи линейного программирования выделяют две задачи, эквивалентные исходной: это так называемая **каноническая** задача, когда требуется найти минимум линейной функции $J(\cdot) = (c, \cdot)$ на множестве

$$u \in \mathcal{U} = \{u \in \mathcal{R}_n,\ u \geq 0,\ Au = b\}, \tag{3}$$

и **основная** задача, когда множество, на котором минимизируется линейная функция $J(\cdot)$, задано в виде

$$u \in \mathcal{U} = \{u \in \mathcal{R}_n,\ u \geq 0,\ Au \leq b\}. \tag{4}$$

Здесь b и c — заданные векторы, $c \in \mathcal{R}_n$, $b \in \mathcal{R}_m$, A — невырожденная матрица размера $m \times n$[2]. Поясним, в каком смысле можно считать эти задачи эквивалентными.

Заменим равенства $Au = b$ равносильной системой неравенств

$$\begin{cases} Au \leq b, \\ Au \geq b. \end{cases}$$

Теперь множество (3) запишется в виде $\mathcal{U} = \{u \in \mathcal{R}_n,\ u \geq 0,\ Au \leq b;\ -Au \leq -b\}$. С другой стороны, для множества (4) введем дополнительные переменные $v = (v_1, \ldots, v_m)$, где $v = b - Au$; согласно (4), $v \geq 0$. Теперь общая задача запишется как следующая каноническая задача линейного программирования в пространстве $\mathcal{R}_{n+m}$:

$$(d, z) \to \inf_{z \in Z \subset \mathcal{R}_{n+m}},$$

$$Z = \{z = (u_1, \ldots, u_n, v_1, \ldots, v_m) \in \mathcal{R}_{n+m},\ z \geq 0,\ \ Cz = Au + v = b\},$$

[2]Матрица A невырождена, если $Ax = 0$ влечет $x = 0$.

где $d = (c_1, \ldots, c_n, 0, \ldots, 0) \in \mathcal{R}_{n+m}$, C — блочная матрица размера $(n+m) \times m$, равная $C = (A \;\; I)$.

Использованный здесь прием интересен с теоретической точки зрения, так как вместо общей задачи линейного программирования позволяет исследовать только основную или каноническую задачу, применение же его на практике приводит к существенному увеличению размерности задачи и не всегда оправдано.

Геометрическая интерпретация задачи линейного программирования и ее обобщение. Пусть для наглядности $n = 2$ и требуется найти минимум функции $J(u) = c_1u_1 + c_2u_2$ на множестве

$$\mathcal{U} = \{u \in \mathcal{R}_2 : \; u_1 \geq 0, \; u_2 \geq 0, \;\; a_{i1}u_1 + a_{i2}u_2 \leq b_i, \;\; i = 1, \ldots, m\}.$$

Обозначим $\mathcal{U}_0 = \{u : \; u_1 \geq 0, \, u_2 \geq 0\}$ — положительный квадрант на плоскости $\mathcal{R}_2$, $\mathcal{U}_i = \{u : \; a_{i1}u_1 + a_{i2}u_2 \leq b_i\}$ — полуплоскость, ограниченная прямой $a_{i1}u_1 + a_{i2}u_2 = b_i$, $i = 1, \ldots, m$. Множество, на котором минимизируется линейная функция, есть пересечение $\mathcal{U} = \mathcal{U}_0 \cap \mathcal{U}_1 \cap \ldots \cap U_m$. Рассмотрим несколько случаев. 1. $\mathcal{U} = \emptyset$. В этом тривиальном случае множество точек минимума функции пусто. 2. Если $\mathcal{U} \neq \emptyset$, то $\mathcal{U}$ — подмножество плоскости $\mathcal{R}_2$, граница которого — ломанная, состоящая из отрезков осей Ou_1 и Ou_2 и прямых $a_{i1}u_1 + a_{i2}u_2 = b_i$, $i = 1, \ldots, m$. Это множество может быть как ограниченным, так и неограниченным. Если оно ограничено, то это — выпуклый многоугольник.

Функция $J(\cdot)$ принимает одно и то же значение $J(u) = (c, u)$, $u \in \mathcal{R}_2$, на прямых вида $c_1u_1 + c_2u_2 = \alpha$, $-\infty < \alpha < \infty$. Эти условия задают семейство прямых — линий уровня функции $J(\cdot)$, каждая из которых ортогональна вектору $c \in \mathcal{R}_2$ и удалена от начала координат на расстояние α. При изменении α от $-\infty$ до ∞ такие прямые "зачертят"всю плоскость $\mathcal{R}_2$, при этом вектор c задает направление, в котором надо передвигать прямую, чтобы увеличить значение функции $J(\cdot)$. Если $\mathcal{U}$ — ограниченное множество, то при изменении α от $-\infty$ до ∞ найдется такая прямая из семейства $c_1u_1 + c_2u_2 = \alpha$, которая впервые коснется многоугольника $\mathcal{U}$ при некотором $\alpha = \alpha_0$. Это касание обязательно произойдет в вершине многоугольника (даже если касание происходит целой стороной многоугольника — при этом есть даже две вершины, в которых происходит касание). Если множество $\mathcal{U}$ неограничено, то в зависимости от его расположения на плоскости $\mathcal{R}_2$ относительно вектора $c \in \mathcal{R}_2$ возможно и $J_* = -\infty$, и $J_* > -\infty$, причем в последнем случае первое касание прямой $c_1u_1 + c_2u_2 = \alpha$ с множеством $\mathcal{U}$ по-прежнему происходит в точке, где пересекаются прямые, ограничивающие множество $\mathcal{U}$ — такая точка называется угловой.

Отсюда можно заметить, что задача линейного программирования может не иметь решений, может иметь единственное решение или целое множество решений; причем если $J_* > -\infty$, то решение достигается в угловой точке множества $\mathcal{U}$.

Дадим определение угловой точки и приведем результат, обобщающий наши геометрические рассуждения.

Определение. Точка $v \in \mathcal{U}$ называется угловой точкой (или вершиной, или крайней точкой, или экстремальной точкой) множества $\mathcal{U}$, если представление

$$v = \alpha v_1 + (1-\alpha)v_2$$

при $v_1, v_2 \in \mathcal{U}$ и $0 < \alpha < 1$ возможно лишь при $v_1 = v_2$. Иначе говоря, v — угловая точка, если она не является внутренней точкой никакого отрезка, целиком содержащегося в $\mathcal{U}$.

Утверждение. Пусть $\mathcal{U} = \{n \geq 0,\ Au = b\}$, матрица A невырождена и ее ранг равен r. Для того, чтобы $v = (v_1, \ldots, v_n) \in \mathcal{U}$ была угловой точкой множества $\mathcal{U}$, необходимо и достаточно, чтобы существовали индексы $j_1, \ldots, j_r$, $1 \leq j_k \leq n$, $k = 1, \ldots, r$, такие, что

$$A_{j_1} v_{j_1} + A_{j_2} v_{j_2} + \ldots + A_{j_r} v_{j_r} = b, \qquad (x)$$

и $v_l = 0$ при всех $l \neq j_k$, $k = 1, \ldots, r$, где $A_{j_1}, A_{j_2}, \ldots, A_{j_r}$ — линейно независимые столбцы матрицы A.

Поясним это утверждение. Имеет смысл рассматривать случай $r < n$, так как иначе (при $r = n$) уравнение $Au = b$ либо имеет единственное решение, либо не имеет решений, а значит, при $r = n$ задача линейного программирования является тривиальной. Если $r < n$, $m > n$, то среди уравнений системы $Au = b$ имеются линейно зависимые, и удаление их из условий $Au = b$, определяющих множество $\mathcal{U}$, не изменит это множество. Далее будем считать, что $m = n$.

Определение. Система векторов $A_{j_1}, A_{j_2}, \ldots, A_{j_r}$, входящих в равенство (x), назовем базисом угловой точки v, а соответствующие им переменные $v_{j_1}, v_{j_2}, \ldots, v_{j_r}$ — базисными координатами угловой точки v. Если среди базисных координат $v_{j_1}, v_{j_2}, \ldots, v_{j_r}$ хотя бы одна равна нулю, то такая угловая точка называется вырожденной.

Если угловая точка невырождена, то она обладает единственным базисом $A_{j_1}, A_{j_2}, \ldots, A_{j_r}$, а если вырождена, то может обладать несколькими базисами.

Пример. Пусть $\mathcal{U} = \{(u_1, u_2, u_3, u_4) \in \mathcal{R}_4 : u_i \geq 0,\ i = 1, \ldots, 4,\ u_1 + u_2 + 3u_3 + u_4 = 3,\ u_1 - u_2 + u_3 + 2u_4 = 1\}$. Тогда

$$A = \begin{pmatrix} 1 & 1 & 3 & 1 \\ 1 & -1 & 1 & 2 \end{pmatrix}, \qquad b = \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$

Ранг матрицы A равен двум. Найдем все угловые точки. Уравнение

$$A_1 v_1 + A_2 v_2 = b$$

разрешимо единственным образом: $v_1 = 2$, $v_2 = 1$. Значит, точка $u^1 = (2, 1, 0, 0)$ — угловая.

Уравнение

$$A_2 v_2 + A_4 v_4 = b$$

разрешимо единственным образом и имеет решение $v_2 = \frac{5}{3}$, $v_4 = \frac{4}{3}$. Следующей угловой точкой является $u^2 = (0, \frac{5}{3}, 0, \frac{4}{3})$.

Из уравнения

$$A_1 v_1 + A_3 v_3 = b$$

получаем $v_1 = 0$, $v_3 = 1$. Значит, угловой точкой является точка $u^3 = (0, 0, 1, 0)$, и ей соответствуют базисы (A_1, A_3), (A_2, A_3) и (A_3, A_4).

Наконец, уравнение

$$A_1 v_1 + A_4 v_4 = b$$

имеет решение $v_1 = 5$, $v_4 = -2$, и точка $u^4 = (5, 0, 0, -2)$ не принадлежит области $\mathcal{U}$, так как одна из ее координат отрицательна.

Таким образом, имеется три угловых точки, подозрительные на экстремум. Вычислив значения функции $J(\cdot)$ в точках u_i, $i = 1, 2, 3$ и выбрав из них наименьшее и соответствующую ей точку, получим решение задачи линейного программирования.

Алгоритм минимизации перебором всех угловых точек весьма трудоемок, так как при большом числе ограничений требуется рассматривать $C_n^r = \frac{n!}{r!(n-r)!}$ уравнений для определения угловых точек. При больших r и n используют направленный перебор угловых точек (даваемый, например, симлекс-методом), и специальный алгоритм выбора начальной точки.

Задачи выпуклого программирования.

Рассмотрим задачу минимизации выпуклой функции $J(\cdot)$, заданной на выпуклом множестве $\mathcal{U}_0 \in \mathcal{R}_n$, в которой требуется найти точную нижнюю грань

$$J_* = \inf\{J(u) | u \in \mathcal{U}\} \tag{3}$$

и хотя бы одну точку u_* из множества

$$\mathcal{U}_* = \{u \in \mathcal{U} : \ J(u) = J_*\},$$

где множество $\mathcal{U}$ задано следующими ограничениями

$$\mathcal{U} = \{u \in \mathcal{U}_0 \subset \mathcal{R}_n, \ g_i(u) \le 0, \ i = 1, \ldots, m; g_i(u) = 0, \ i = m+1, \ldots, s\}, \tag{4}$$

где $g_1(\cdot), \ldots, g_m(\cdot)$ — определены и выпуклы на $\mathcal{U}_0$, $g_i(u) = (a_i u) - b_i$, $i = m+1, \ldots, s < \infty$.

В этих задачах важную роль играет функция Лагранжа

$$L(u, \lambda) = J(u) + \sum_{i=1}^{s} \lambda_i g_i(u), \tag{5}$$

определенная для всех $u \in \mathcal{U}_0$ и для всех

$$\lambda \in \Lambda_0 = \{\lambda \in \mathcal{R}_s, \ \lambda_1 \ge 0, \ldots, \lambda_m \ge 0\}. \tag{6}$$

Определение. Точка (u_*, λ^*) называется седловой точкой функции Лагранжа $L(\cdot,\cdot)$, если

$$L(u_*, \lambda) \le L(u_*, \lambda^*) \le L(u, \lambda^*). \tag{7}$$

для всех $u \in \mathcal{U}_0$, $\lambda \in \Lambda_0$.

Теорема 1. *Для функции Лагранжа (6) точка (u_*, λ^*) будет седловой тогда и только тогда, когда*

$$L(u_*, \lambda^*) \le L(u, \lambda^*)$$

и

$$\lambda_i^* g_i(u_*) = 0, \quad u_* \in \mathcal{U}, \quad i = 1, \dots, s. \tag{8}$$

Заметим, что определение седловой точки (7) и теорема 1 выполнены независимо от того, выпуклы или нет функции $J(\cdot)$, $g_i(\cdot)$, $i = 1, \dots, m$.

Доказательство. Пусть вначале выполнены неравенства (7). Покажем, что тогда $u_* \in \mathcal{U}$ и выполняется (8). Воспользуемся видом функции Лагранжа (5) и запишем левое неравенство (7) в виде

$$\sum_{i=1}^{s} (\lambda_i^* - \lambda_i) g_i(u_*) \ge 0, \quad \lambda \in \Lambda_0. \tag{9}$$

Выберем вначале $\lambda_j = \lambda_j^* + 1$ для некоторого фиксированного j, $1 \le j \le m$, и $\lambda_j = \lambda_j^*$ для всех $i \ne j$, $i = 1, \dots, s$. Т.к. $\lambda^* \in \Lambda_0$, то и $\lambda \in \Lambda_0$, и (9) влечет $g_j(u_*) \le 0$, $j = 1, \dots, m$.

Если же выбрать $\lambda_j = \lambda_j^* + g_j(u_*)$ для некоторого фиксированного j, $m+1 \le j \le s$, и $\lambda_j = \lambda_j^*$ для всех $i \ne j$, $i = 1, \dots, s$, то (9) влечет $-(g_j(u_*))^2 \ge 0$, т.е. $g_j(u_*)) = 0$ для всех $j = m+1, \dots, s$.

Выберем теперь $\lambda \in \Lambda_0$ так, чтобы $\lambda_j = 0$ для некоторого фиксированного j, $1 \le j \le m$, и $\lambda_j = \lambda_j^*$ для всех $i \ne j$, $i = 1, \dots, s$. Тогда, согласно (9), $\lambda_i^* g_i(u_*) \ge 0$, а $\lambda_i^* \ge 0$, $g_i(u_*) \le 0$, поэтому $\lambda_i^* g_i(u_*) = 0$, $i = 1, \dots, m$.

Пусть теперь выполнено (8). Покажем, что (u_*, λ^*) — седловая точка функции Лагранжа (5), т.е. выполнено левое неравенство (7). Согласно (8), $u_* \in \mathcal{U}$, т.е. $g_i(u_*) \le 0$, $i = 1, \dots, m$, $g_i(u_*) = 0$, $i = m+1, \dots, s$, и

$$(\lambda_i^* - \lambda_i) g_i(u_*) = -\lambda_i g_i(u_*) \ge 0,$$

т.к. для $i = m+1, \dots, s$ $g_i(u_*) = 0$, а для $i = 1, \dots, m$ $\lambda_i^* \ge 0$, а $g_i(u_*) \le 0$. Отсюда следует (9), а значит и левые неравенства в (7).

Теорема 2. *Пусть $(u_*, \lambda^*) \in \mathcal{U}_0 \otimes \Lambda_0$ — седловая точка функции Лагранжа. Тогда $u_* \in \mathcal{U}_*$, $J_* = L(u_*, \lambda^*) = J(u_*)$, т.е. u_* — решение задачи (3), (4).*

Доказательство. Согласно условиям (8), $u_* \in \mathcal{U}$ и $L(u_*, \lambda^*) = J(u_*)$. Поэтому правая группа неравенств (7) примет вид

$$J(u_*) \le L(u, \lambda^*) = J(u) + \sum_{i=1}^{s} \lambda_i^* g_i(u), \quad u \in \mathcal{U}_0,$$

и в частности, оно выполнено для $u \in \mathcal{U}$. А т.к. для $u \in \mathcal{U}$ $g_i(u) \le 0$, $i = 1, \dots, m$ и $g_i(u) = 0$, $i = m+1, \dots, s$, а $\lambda_i^* \ge 0$ для $i = 1, \dots, m$, то для $u \in \mathcal{U}$ $\sum\limits_{i=1}^{s} \lambda_i^* g_i(u) \le 0$. Поэтому для $u \in \mathcal{U}$

$$J(u_*) \le L(u, \lambda^*) \le J(u) \quad \text{для всех} \quad u \in \mathcal{U},$$

т.е. $u_* \in \mathcal{U}_*$.

Однако в общей ситуации вопрос о существовании седловой точки функции Лагранжа при $\mathcal{U}_* \ne \emptyset$ остается открытым, и значит, не всякое решение задачи (3)-(4) можно получить путем отыскания седловой точки функции Лагранжа.

Теоремы, в которых устанвливается существование седловой точки функции Лагранжа, называются теоремами Куна-Таккера.

Одним из условий теорем Куна-Таккера является условие регулярности множества $\mathcal{U}$. Остановимся на случае, когда множество $\mathcal{U}$ задано только неравенствами:

$$\mathcal{U} = \{u \in \mathcal{U}_0,\ g_i(u) \le 0,\ i = 1, \dots, m\}. \tag{10}$$

Множество (10) (и каждое ограничение в (10)) называется регулярным, если $\mathcal{U}_0$ выпукло, $g_i(\cdot)$, $i = 1, \dots, m$ выпуклы на $\mathcal{U}_0$ и для любого $i = 1, \dots, m$ существует точка $u_i \in \mathcal{U}$, в которой $g_i(u_i) < 0$.

Заметим, что если $\mathcal{U}_0$ регулярно, то существует и точка $\widetilde{u} \in \mathcal{U}_0$, для которой

$$g_i(\widetilde{u}) < 0, \quad i = 1, \dots, m. \tag{11}$$

Действительно, пусть $g_i(u_i) < 0$, $u_i \in \mathcal{U}$; тогда для $\widetilde{u} = \sum\limits_{i=1}^{m} \alpha_i u_i$, для чисел $\alpha_i \ge 0$, $i = 1, \dots, m$, $\sum\limits_{i=1}^{m} \alpha_i = 1$ выполнено: $\widetilde{u} \in \mathcal{U}$ и $g_i(\widetilde{u}) \le \sum\limits_{i=1}^{m} \alpha_i g_i(u_i) < 0$ в силу выпуклости множества $\mathcal{U}$ и функций $g_i(\cdot)$, $i = 1, \dots, m$. Условие (11) называются условием Слейтера.

Теорема Куна-Таккера. *Пусть множество $\mathcal{U}_0$ выпукло, функции $J(\cdot)$, $g_i(\cdot)$, $i = 1, \dots, m$ выпуклы на $\mathcal{U}_0$ и множество $\mathcal{U}$ в (10) регулярно. Если множество $\mathcal{U}_*$ точек минимума функции $J(\cdot)$ на $\mathcal{U}$ не пусто, то для любой точки $u_* \in \mathcal{U}_*$ существуют множители Лагранжа*

$$\lambda^* = (\lambda_1^*, \dots, \lambda_m^*) \in \Lambda_0 = \{\lambda \in \mathcal{R}_m,\ \lambda_i \ge 0,\ i = 1, \dots, m < \infty\},$$

такие, что (u_*, λ^*) является седловой точкой функции Лагранжа на множестве $\mathcal{U}_0 \otimes \Lambda_0$.

Если $g_i(\cdot)$ — линейные функции вида $g_i(\cdot) = (a_i, \cdot) + b_i$, $i = 1, \dots, m$, то утверждение теоремы верно и без требования регулярности ограничений.

Итак, если (u_*, λ^*) — седловая точка функции Лагранжа, то u_* — решение задачи (3)–(4) и без предположения о выпуклости функций $J(\cdot)$, $g_1(\cdot)$,…,$g_m(\cdot)$.

Условие выпуклости этих функций и регулярности множества (10) гарантирует, что u_* — решение задачи (3), (10) тогда и только тогда, когда (u_*, λ^*) — седловая точка функции Лагранжа при некотором $\lambda^* \in \Lambda_0$. В этом случае решение задачи (3), (11) может быть найдено из условий

$$\begin{array}{l} L(u_*, \lambda^*) \le L(u, \lambda^*), \quad u \in \mathcal{U}_0, \quad \lambda^* = (\lambda_1^*, \ldots, \lambda_m^*), \lambda^* \ge 0, \\ \lambda_i^* g_i(u_*) = 0, \quad i = 1, \ldots, m, \quad u_* \in \mathcal{U}. \end{array} \tag{12}$$

Так как первое условие в (12) означает, что u_* — точка минимума функции Лагранжа $L(u, \lambda^*)$, $u \in \mathcal{U}_0$, то в силу дифференцируемости функции $L(u, \lambda^*)$ по $u \in \mathcal{U}_0$ его можно записать в виде

$$(L'_u(u_*, \lambda^*), u - u_*) \ge 0. \tag{13}$$

Если, кроме того, $\mathcal{U}_0 = \mathcal{R}_n$, то (13) примет вид

$$L'_u(u_*, \lambda^*) = J'(u_*) + \sum_{i=1}^{m} \lambda_i^* g_i'(u_*) = 0.$$

Пример. Пусть требуется найти минимум функции $J(u) = \|Au - y\|^2$ при условии $\|D(u - u_0)\|^2 \le \varepsilon$. Здесь $u \in \mathcal{R}_N$, $y \in \mathcal{R}_n$, $A \in (\mathcal{R}_N \to \mathcal{R}_n)$, $D \in (\mathcal{R}_N \to \mathcal{R}_N)$ — линейные операторы, причем оператор D невырожден, т.е. $Du = 0$ тогда и только тогда, когда $u = 0$. Эту вариационную задачу можно записать в виде

$$\inf\{\|Au - y\|^2 \mid \|D(u - u_0)\|^2 \le \varepsilon\}.$$

Заменим переменные, обозначив $v = u - u_0$, $z = y - Au_0$; в новых переменных вариационная задача примет вид

$$\inf\{\|Av - z\|^2 \mid \|Dv\|^2 \le \varepsilon\}. \tag{14}$$

Запишем функцию Лагранжа в виде

$$L(v, \lambda) = \|Av - z\|^2 + \lambda(\|Dv\|^2 - \varepsilon), v \in \mathcal{R}_N, \quad \lambda \ge 0.$$

Задача (14) является задачей выпуклого программирования, следовательно, при $\varepsilon > 0$ условие регулярности выполнено. Необходимые и достаточные условия минимума имеют вид:

$$\frac{\partial L(v, \lambda)}{\partial v} = A^*(Av - z) + \lambda D^* Dv = 0,$$

$$\lambda(\|Dv\|^2 - \varepsilon) = 0,$$

$$\lambda \ge 0, \quad \|Dv\|^2 \ge \varepsilon.$$

Эти условия называются вариационными. При $\lambda > 0$ первое уравнение в вариационных условиях имеет решение

$$v(\lambda) = (A^*A + \lambda D^*D)^{-1}A^*z.$$

Проверим, при каких $\varepsilon > 0$ найдется такое $\lambda > 0$, при которых выполняется вариационное условие

$$\|Dv(\lambda)\|^2 = \varepsilon.$$

Для этого для положительно определенного самосопряженного оператора $D^*D \in (\mathcal{R}_N \to \mathcal{R}_N)$ введем обозначение $Q = D^*D = Q^{1/2}Q^{1/2}$, где $Q^{1/2} \in (\mathcal{R}_N \to \mathcal{R}_N)$ — квадратный корень из Q, т.е. положительно определенный самосопряженный оператор, квадрат которого равен Q; $Q^{-1/2}$ — оператор, обратный к $Q^{1/2}$. Запишем

$$(A^*A + \lambda D^*D)^{-1}A^* = (Q^{1/2}Q^{-1/2}A^*AQ^{-1/2}Q^{1/2} + \lambda Q^{1/2}Q^{1/2})^{-1}A^* =$$

$$= Q^{-1/2}(Q^{-1/2}A^*AQ^{-1/2} + \lambda I)^{-1}Q^{-1/2}A^* = Q^{-1/2}Q^{-1/2}A^*(AQ^{-1}A^* + \lambda I)^{-1},$$

обозначим $B = AQ^{-1/2}$ и найдем собственные значения и соответствующие им собственные векторы самосопряженного оператора $BB^* \in (\mathcal{R}_n \to \mathcal{R}_n)$:

$$BB^*e_\mu = \beta_\mu^2 e_\mu, \quad \mu = 1, \ldots, n.$$

Векторы $\{e_\mu\}$ выберем так, чтобы они составляли полную ортонормированную систему в $\mathcal{R}_n$ (базис в $\mathcal{R}_n$); этому базису соответствует оргонормированный базис $\{h_\mu\} \subset \mathcal{R}_N$:

$$B^*e_\mu = \beta_\mu h_\mu, \quad \mu = 1, \ldots, \min\{n, N\}.$$

Проаналиируем поведение функции $S(\lambda) = \|Dv(\lambda)\|^2$ при $\lambda > 0$, для чего запишем эту функцию с использованием базисов $\{h_\mu\} \subset \mathcal{R}_N$ и $\{e_\mu\} \subset \mathcal{R}_n$:

$$S(\lambda) = (D^*Dv(\lambda), v(\lambda)) = \|Q^{1/2}v(\lambda)\|^2 = \left\| B^*(BB^* + \lambda I)^{-1} \sum_{\mu=1}^{n} (y, e_\mu)e_\mu \right\|^2 =$$

$$= \left\| \sum_{\mu=1}^{N} \frac{\beta_\mu (y, e_\mu)}{\beta_\mu^2 + \lambda} h_\mu \right\|^2 = \sum_{\mu=1}^{N} \frac{\beta_\mu^2 (y, e_\mu)^2}{(\beta_\mu^2 + \lambda)^2}.$$

Таким образом функция $S(\lambda)$ является суммой конечного числа монотонно убывающих функций, и следовательно, сама монотонно убывает на множестве $\lambda \geq 0$ от значения

$$\varepsilon_0 = \lim_{\lambda \to 0} S(\lambda) = \sum_{\beta_\mu \neq 0} \frac{(y, e_\mu)^2}{\beta_\mu^2}$$

при $\lambda = 0$ до нуля при $\lambda \to \infty$.

Итак, при $0 < \varepsilon < \varepsilon_0$ все вариационные условия выполняются при

$$v = v(\lambda_\varepsilon) = (A^*A + \lambda_\varepsilon D^*D)^{-1}A^*z,$$

где $\lambda_\varepsilon > 0$ — единственный корень уравнения

$$\|Dv(\lambda_\varepsilon)\|^2 = \varepsilon.$$

При $\varepsilon \geq \varepsilon_0$ вариационные условия не могут быть выполнены при $\lambda > 0$, поэтому следует положить $\lambda = 0$; взяв предел $v(0) = \lim\limits_{\lambda\to 0} v(\lambda) = Q^{-1/2}B^-z$, заметим, что $(v(0), 0)$ является седловой точкой функции Лагранжа. Действительно, $\dfrac{\partial L}{\partial v}|_{v=v(0)} = 0$, поскольку условие $A^*Av = A^*z$, получающееся из равенства нулю производной функции Лагранжа по v при $\lambda = 0$, эквивалентно уравнению

$$Q^{-1/2}A^*AQ^{-1/2}Q^{1/2}v = Q^{-1/2}A^*z,$$

или, с учетом введенных обозначений,

$$B^*BQ^{1/2}v = B^*z,$$

откуда и получим искомое решение:

$$v = Q^{-1/2}B^-z = v(0).$$

Неравенство

$$L(v(0), 0) \geq L(v(0), \lambda), \quad \lambda \geq 0,$$

дающее второе условие для седловой точки функции Лагранжа, в данном случае записывается в виде

$$0 \geq \lambda(S(0) - \varepsilon),$$

выполняющееся в силу условия $\lambda = 0$.

Окончательно, решение исходной вариационной задачи запишется в виде

$$u = u_0 + \begin{cases} (A^*A + \lambda_\varepsilon D^*D)^{-1}A^*(y - Au_0) & \text{при } 0 < \varepsilon < \varepsilon_0, \\ \lim\limits_{\lambda\to 0}(A^*A + \lambda_\varepsilon D^*D)^{-1}A^*(y - Au_0) & \text{при } \varepsilon \geq \varepsilon_0, \end{cases}$$

где $\varepsilon_0 = \lim\limits_{\lambda\to 0} \|D(A^*A + \lambda D^*D)^{-1}A^*(y - Au_0)\|^2$, а λ_ε — единственное решение уравнения

$$\|D(A^*A + \lambda D^*D)^{-1}A^*(y - Au_0)\|^2 = \varepsilon.$$

Часть 3. Задачи на минимакс.

С задачами на минимакс мы встречаемся, например, при минимизации максимально возможной погрешности оценивания параметров изучаемого объекта, или при минимизации максимальных потерь при выборе стратегии поведения в той или иной ситуации. По существу, минимакс является одним из способов оптимального выбора параметров.

Пусть $J(\cdot,\cdot)$ — функция, заданная на $\mathcal{U}_1 \otimes \mathcal{U}_2$, где $\mathcal{U}_1$ — выпуклое замкнутое множество евклидова пространства $\mathcal{R}_1$, а $\mathcal{U}_2$ — ограниченное замкнутое множество евклидова пространства $\mathcal{R}_2$. В задаче на минимакс требуется найти минимум функции $\varphi(\cdot) = \max\limits_{u_2 \in \mathcal{U}_2} J(\cdot, u_2)$ на множестве $\mathcal{U}_1$.

Если $J(\cdot,\cdot)$ линейна по первому аргументу при фиксированном втором, то говорят о линейных задачах на минимакс, в противном случае речь идет о нелинейном минимаксе.

Глава 5. Чебышевская интерполяция.

Пусть задана таблица значений функции $y(t_k) = y_k$, $k = 0, 1, \ldots, N$, где $t_0 < t_1 < \ldots < t_N$ — заданные значения аргумента функции $y(\cdot)$, которые будем называть узлами сетки. Тогда любой полином $P_n(\mathbf{a}, t) = \sum\limits_{j=0}^{n} a_j t^j$, $t \in [a, b]$, $a \le t_0$, $b \ge t_N$, $\mathbf{a} = (a_0, a_1, \ldots, a_N) \in \mathcal{R}_{n+1}$, степень которого n не выше, чем N, имеет по отношению к заданному набору чисел естественную характеристику — величину максимального уклонения, равную

$$\max_{k \in \{0,1,\ldots,N\}} |y_k - P_n(\mathbf{a}, t_k)|.$$

Задача состоит в поиске такого полинома $P_n(\mathbf{a}, t)$, для которого величина максимального уклонения минимальна. Формально для этого требуется выбрать коэффициенты полинома $P_n(\mathbf{a}, t)$ из задачи на минимакс

$$\rho = \inf_{\mathbf{a} \in \mathcal{R}_{n+1}} \max_{k \in \{0,1,\ldots,N\}} |y_k - P_n(\mathbf{a}, t_k)|$$

при фиксированном n. Полином $P_n(\mathbf{a}^*, t)$, на котором достигается минимум максимального уклонения, то есть для которого $\rho = \max\limits_{k \in \{0,1,\ldots,N\}} |y_k - P_n(\mathbf{a}^*, t_k)|$, называется полиномом наилучшего приближения.

Интерполяция полиномом степени $n = N - 1$. Если $N = n$, то $\rho = 0$, и полином наилучшего приближения — это обыкновенный интерполяционный полином. Первой нетривиальной ситуацией является случай, когда $n = N - 1$. Он приводит к так называемой чебышевской интерполяции, и на нем основан и более общий подход, когда $n < N$.

Теорема 1. Пусть $\rho = \inf\limits_{\mathbf{a}\in\mathcal{R}_{n+1}} \max\limits_{k\in\{0,1,\ldots,N\}} |y_k - P_n(\mathbf{a}, t_k)|$. Полином наилучшего приближения существует и единственен. Для того, чтобы полином $P_n(\mathbf{a}^*, t)$ был полиномом наилучшего приближения, необходимо и достаточно, чтобы при некотором h выполнялось равенство

$$(-1)^k h + P_n(\mathbf{a}^*, t_k) = y_k \quad \forall k = 0, 1, \ldots, n+1. \tag{1}$$

Докажем лишь достаточность. Пусть для $P_n(\mathbf{a}^*, t)$ при некотором h справедливо равенство (1). Покажем, что $P_n(\mathbf{a}^*, t)$ — полином наилучшего приближения. Предположим, что это не так, то есть

$$\rho < \max_{k\in\{0,1,\ldots,N\}} |y_k - P_n(\mathbf{a}, t_k)| \equiv |h|.$$

Тогда в силу определения ρ существует полином $P_n(\mathbf{a}_1, t)$, такой, что

$$\max_{k\in\{0,1,\ldots,N\}} |y_k - P_n(\mathbf{a}_1, t_k)| < |h|. \tag{2}$$

Рассмотрим полином n-ной степени $Q_n(t) = P_n(\mathbf{a}_1, t) - P_n(\mathbf{a}^*, t)$; в узлах $t_0, t_1, \ldots, t_N$ для него выполнены равенства

$$Q_n(t_k) = (P_n(\mathbf{a}_1, t_k) - y_k) + (y_k - P_n(\mathbf{a}^*, t)) = (-1)^k h + (y_k - P_n(\mathbf{a}_1, t)).$$

В силу (2) числа $Q_n(t_k)$ имеют тот же знак, что и $(-1)^k h$, $k = 0, 1, \ldots, N$, что означает, что полином n-ной степени последовательно меняет свой знак в $N+1 = n+2$ точках; так как полином — непрерывная функция, то между узлами, в которых он принимает значения разных знаков, полином имеет корень. Таким образом, мы приходим к выводу, что полином n-ной степени $Q_n(\cdot)$ имеет $n+1$ корень, что возможно лишь при $Q_n(t) \equiv 0$. Итак, достаточность доказана, и попутно показано, что $\rho = |h|$.

Определение. Полином, удовлетворяющий соотношению (1), называется чебышевским интерполяционным полиномом.

Установим явный вид h и алгоритм построения чебышевского интерполяционного многочлена. Обозначим $\chi_k = \chi(t_k) = (-1)^k$, $k = 0, 1, \ldots, n+1$, и для произвольного набора чисел $y(t_0), y(t_1), \ldots, y(t_{n+1})$ определим разделенные разности по формулам

$$y[t_s, t_{s+1}] = \frac{y(t_{s+1}) - y(t_s)}{t_{s+1} - t_s,},$$

$$y[t_s, t_{s+1}, t_{s+2}] = \frac{y[t_{s+1}, t_{s+2}] - y[t_s, t_{s+1}]}{t_{s+2} - t_s,},$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$y[t_s, t_{s+1}, \ldots, t_{s+l}] = \frac{y[t_{s+1}, \ldots, t_{s+l}] - y[t_s, t_{s+l-1}]}{t_{s+l} - t_s,},$$

.....

Тогда

$$h = \frac{y[t_0, t_1, \ldots, t_N]}{\chi[t_0, t_1, \ldots, t_N]}. \tag{3}$$

Действительно, для любого h существует полином $L_{n+1}(t)$ степени не выше $n+1$, однозначно определенный соотношениями

$$L_{n+1}(t_k) = w_k, \qquad k = 0, 1, \ldots, n+1,$$

где $w_k = y_k - h\chi_k$, $k = 0, 1, \ldots, n+1$, его можно записать через разделенные разности в виде

$$L_{n+1}(t) = w_0 + \sum_{k=1}^{n+1} w[t_0, t_1, \ldots, t_k](t - t_0)(t - t_1)\ldots(t - t_{k-1}). \tag{4}$$

Коэффициент при старшей степени t^{n+1} равен $w[t_0, t_1, \ldots, t_{n+1}] = y[t_0, t_1, \ldots, t_{n+1}] - h\chi[t_0, t_1, \ldots, t_{n+1}]$, и в силу (1) он обращается в нуль. Так что при выполнении равенств (1) полином имеет степень не выше n и тем самым, является чебышевским интерполяционным полиномом наилучшего приближения, так как для него выполнены достаточные условия теоремы 1.

Итак, для построения чебышевского интерполяционного полинома необходимо выполнить следующие действия.

1. Построить разделенные разности для функций $y(t)$ и $\chi(t)$.
2. Найти h по формуле (3).
3. Вычислить полином наилучшего приближения по формуле (4).

Общий случай интерполяции, $n < N$. Пусть теперь, по прежнему, задана таблица из N значений $\{y_k\}$ функции $y(t)$, $k = 0, 1, \ldots, N$ в узлах $t_0 < t_1 < \ldots < t_N$, для полинома $P_n(\mathbf{a}, t)$ степени n, $N > n+1$, определим его максимальное уклонение от $y(t)$ величиной

$$\max_{k\in\{0,1,\ldots,N\}} |y_k - P_n(\mathbf{a}, t_k)|,$$

и обозначим

$$\rho = \inf_{\mathbf{a}\in\mathcal{R}_{n+1}} \max_{k\in\{0,1,\ldots,N\}} |y_k - P_n(\mathbf{a}, t_k)|.$$

Задача состоит в поиске полинома $P_n(\mathbf{a}^*, t)$, для которого

$$\rho = \max_{k\in\{0,1,\ldots,N\}} |y_k - P_n(\mathbf{a}^*, t_k)|.$$

Определение. Базисом σ назовем любое $(n+2)$-точечное подмножество узлов: $\sigma = \{t_{i_0} < t_{i_1} < \ldots < t_{i_{n+1}}\}$.

Множество базисов $\{\sigma\}$ обозначим символом Σ.

Согласно предыдущей теореме 1, для любого базиса σ можно определить чебышевский интерполяционный полином. Пусть $\rho(\sigma)$ — наилучшее приближение на базисе σ:

$$\rho(\sigma) = \max_{k\in\{0,1,\ldots,n+1\}} |y_{i_k} - P_n(\mathbf{a}(\sigma), t_{i_k})|.$$

Обозначим

$$\varphi(a(\sigma)) = \max_{k\in\{0,1,\ldots,N\}} |y_k - P_n(\mathbf{a}(\sigma), t_k)|,$$

то есть $\varphi(a(\sigma))$ — максимальное уклонение полинома $P_n(\mathbf{a}(\sigma), t)$ от функции $y(t)$ на всем множестве узлов $t_0, t_1, \ldots, t_N$. В силу определений, для любого $\sigma \in \Sigma$ выполняется неравенство $\varphi(a(\sigma)) \geq \rho(\sigma)$.

Утверждение 1. Если для некоторого базиса $\sigma^* \in \Sigma$, $\sigma^* = \{t_{i_0}, t_{i_1}, \ldots, t_{i_{n+1}}\}$ выполняется равенство

$$\varphi(a(\sigma^*)) = \rho(\sigma^*),$$

то

$$\varphi(a(\sigma^*)) \equiv \max_{k\in\{0,1,\ldots,N\}} |y_k - P_n(a(\sigma^*), t_k)| = \rho,$$

то есть $P_n(a(\sigma^*), t_k)$ — полином наилучшего приближения.

Обратимся теперь к тем базисам $\sigma \in \Sigma$, для которых $\varphi(a(\sigma)) > \rho(\sigma)$. Если это так, то можно так изменить базис, что для нового базиса $\sigma_1 \in \Sigma$ выполняется неравенство $\rho(\sigma_1) > \rho(\sigma)$. Для этого, например, достаточно выбрать узел t_{k_0}, в котором $|y_{k_0} - P_n(a(\sigma), t_{k_0})| = \varphi(a(\sigma))$, и заменить его на точку из набора $\{t_0, t_1, \ldots, t_N\}$, в которой отклонение $y(t) - P_n(a(\sigma), t)$ по модулю не меньше, чем $\rho(\sigma)$ и имеет тот же знак.

Утверждение 2. Если для некоторого базиса $\sigma \in \Sigma$ выполняется неравенство

$$\varphi(a(\sigma)) > \rho(\sigma),$$

то существует некоторый базис $\sigma_1 \in \Sigma$, для которого $\rho(\sigma_1) > \rho(\sigma)$.

Определение. Базис $\sigma^* \in \Sigma$, для которого

$$\rho(\sigma^*) = \overline{\rho} = \max_{\sigma\in\Sigma} \rho(\sigma),$$

называется экстремальным базисом.

Экстремальный базис может быть построен конструктивно за конечное число шагов, так как множество Σ конечно.

Теорема 2. При $N \geq n + 1$ полином наилучшего приближения существует и единственен. Для того, чтобы $P_n(\mathbf{a}^*, t)$ был полиномом наилучшего приближения, необходимо и достаточно, чтобы он осуществлял чебышевскую интерполяцию на некотором экстремальном базисе $\sigma^* \in \Sigma$, то есть чтобы

$$P_n(\mathbf{a}^*, t) = P_n(\mathbf{a}(\sigma^*), t),$$

и при этом

$$\inf_{\mathbf{a}\in\mathcal{R}_{n+1}} \max_{k\in\{0,1,\ldots,N\}} |y_k - P_n(\mathbf{a}^*, t_k)| = \max_{\sigma\in\Sigma} \rho(\sigma).$$

Таким образом, для построения чебышевского интерполяционного полинома требуется перебрать базисы $\sigma \in \Sigma$ и остановиться на том, для которого $\rho(\sigma)$ максимально.

Приближение непрерывных функций алгебраическими многочленами. Рассмотренный в предыдущих пунктах подход может быть распространен и на случай, когда непрерывная функция $y(t)$ задана не в конечном числе узлов, а во всех точках t интервала $[c, d]$. В этом случае тоже следует найти экстремальный базис, однако теперь перебор невозможен, и речь может идти о последовательности базисов, максимизирующих функцию $\rho(\sigma)$. Сформулируем основной результат.

Пусть $y(\cdot)$ — непрерывная функция, заданная на отрезке $[c, d]$,

$$\rho_n(y; c, d) = \inf_{\mathbf{a}\in\mathcal{R}_{n+1}} \max_{c\le t\le d} |y(t) - P_n(\mathbf{a}, t)|.$$

Полином $P_n(\mathbf{a}^*, t)$, для которого

$$\max_{c\le t\le d} |y(t) - P_n(\mathbf{a}^*, t)| = \rho_n(y; c, d),$$

называется полиномом наилучшего приближения. Задача состоит в построении полинома наилучшего приближения.

Теорема 3 (П.Л.Чебышева). Полином наилучшего приближения $P_n(\mathbf{a}^*, t)$ функции $y(t)$ на отрезке $[c, d]$ существует и единственен. Для того, чтобы $P_n(\mathbf{a}^*, t)$ был полиномом наилучшего приближения, необходимо и достаточно, чтобы он осуществлял чебышевскую интерполяцию на некотором экстремальном базисе $\sigma^* \in \Sigma$.

Глава 6. Дискретная задача на минимакс.

Постановка задачи Пусть $J_i(\cdot)$, $i = 0, 1, \ldots, N$ — функции, заданные на всем евклидовом пространстве R_n, $n < \infty$. Обозначим

$$J_* = \inf_{u \in \mathcal{R}_n} \max_{i \in \{0,1,\ldots,N\}} J_i(u).$$

Задача на минимакс состоит в нахождении элемента $u_* \in \mathcal{R}_n$, для которого выполнено равенство

$$\max_{i \in \{0,1,\ldots,N\}} J_i(u_*) = J_*.$$

Введем обозначение $J(\cdot) = \max\limits_{i \in \{0,1,\ldots,N\}} J_i(\cdot)$. Так как $J_* = \inf\limits_{u \in \mathcal{R}_n} J(u)$, то изучим свойства функции максимума $J(\cdot)$.

Свойство 1. Если все функции $J_i(\cdot)$, $i = 0, 1, \ldots, N$, непрерывны в точке $u_0 \in \mathcal{R}_n$, то и функция $J(\cdot) = \max\limits_{i \in \{0,1,\ldots,N\}} J_i(\cdot)$ непрерывна в точке u_0.

Для доказательства этого свойства заметим сначала, что для любого набора чисел $F_i^{(1)}$, $F_i^{(2)}$, $i = 0, 1, \ldots, N$, выполнено очевидное неравенство

$$\max_{i \in \{0,1,\ldots,N\}} (F_i^{(1)} + F_i^{(2)}) \le \max_{i \in \{0,1,\ldots,N\}} F_i^{(1)} + \max_{i \in \{0,1,\ldots,N\}} F_i^{(2)}(u).$$

Для $F_i^{(1)} = J_i(u_0) - J_i(u)$, $F_i^{(2)} = J_i(u)$, получим, что в любых фиксированных точках u и u_0 справедливо неравенство

$$\max_{i \in \{0,1,\ldots,N\}} J_i(u_0) \le \max_{i \in \{0,1,\ldots,N\}} J_i(u) + \max_{i \in \{0,1,\ldots,N\}} |J_i(u_0) - J_i(u)|,$$

и аналогично для $F_i^{(2)} = -J_i(u_0) + J_i(u)$, $F_i^{(1)} = J_i(u_0)$,

$$\max_{i \in \{0,1,\ldots,N\}} J_i(u) \le \max_{i \in \{0,1,\ldots,N\}} J_i(u_0) + \max_{i \in \{0,1,\ldots,N\}} |J_i(u_0) - J(u_i)|.$$

а значит,

$$|J(u) - J(u_0)| = |\max_{i \in \{0,1,\ldots,N\}} J_i(u) - \max_{i \in \{0,1,\ldots,N\}} J_i(u_0)| \le \max_{i \in \{0,1,\ldots,N\}} |J_i(u_0) - J_i(u)| \to 0$$

при $u \to u_0$.

Свойство 2. Если все функции $J_i(\cdot)$, $i = 0, 1, \ldots, N$, непрерывны на $\mathcal{R}_n$, и при некотором $v \in \mathcal{R}_n$ лебегово множество

$$\Lambda_{J(v)} = \{u \in \mathcal{R}_n : \ J(u) \le J(v)\}$$

ограничено, то существует точка $u_* \in \mathcal{R}_n$ в которой функция $J(\cdot)$ достигает своего минимального значения.

Это свойство является следствием теоремы Вейерштрасса, приведенной в главе 2.

Свойство 3. Пусть $\mathcal{U} \subseteq \mathcal{R}_n$ — выпуклое множество, и все функции $J_i(\cdot)$, $i = 0, 1, \ldots, N$, выпуклы на $\mathcal{U}$. Тогда функция максимума $J(\cdot)$ выпукла на $\mathcal{U}$.

Это свойство также уже упомянуто в главе 3.

Зафиксируем $u \in \mathcal{R}_n$ и рассмотрим множество индексов

$$R(u) = \{i \in \{0, 1, \ldots, N\} : \ J_i(u) = J(u)\},$$

то есть $R(u)$ — множество тех индексов, при которых функция $J_i(u)$ как функция индексов $i \in \{0, 1, \ldots, N\}$ достигает своего максимального значения, равного $J(u)$.

Утверждение 1. Если все функции $J_i(\cdot)$, $i = 0, 1, \ldots, N$, непрерывны в точке $u_0 \in \mathcal{R}_n$, то найдется такое число $\alpha_0 > 0$, что для любого $\alpha \in [0, \alpha_0]$ и для любого элемента $h \in \mathcal{R}_n$ единичной нормы выполнено равенство

$$J(u_0 + \alpha h) \equiv \max_{i \in \{0,1,\ldots,N\}} J_i(u_0 + \alpha h) = \max_{i \in R(u)} J_i(u_0 + \alpha h).$$

Определение. Функция $J(\cdot)$ называется дифференцируемой в точке $u \in \mathcal{R}_n$ по направлению $h \in \mathcal{R}_n$, $\|h\| = 1$, если существует конечный предел

$$\lim_{\alpha \to 0} \frac{J(u + \alpha h) - J(u)}{\alpha}.$$

Этот предел называется производной $\frac{\partial J(u)}{\partial h}$ функции $J(\cdot)$ в точке $u \in \mathcal{R}_n$ по направлению $h \in \mathcal{R}_n$.

Если $J(\cdot)$ дифференцируема в точке $u \in \mathcal{R}_n$, и $J'(u) \in \mathcal{R}_n$ — ее производная, то ее производная по направлению $h \in \mathcal{R}_n$ равна $\frac{\partial J(u)}{\partial h} = (J'(u), h)$.

Теорема 1. Пусть $J_i(\cdot)$, $i = 0, 1, \ldots, N$, непрерывны и дифференцируемы в некоторой δ-окрестности $O_\delta(u_0) \subset \mathcal{R}_n$ точки $u_0 \in \mathcal{R}_n$. Тогда функция максимума $J(\cdot)$ дифференцируема в точке u_0 по любому направлению $h \in \mathcal{R}_n$, $\|h\| = 1$, причем

$$\frac{\partial J(u_0)}{\partial h} = \max_{i \in R(u)} (J_i'(u_0), h).$$

Доказательство. Для любого индекса $i \in \{0, 1, \ldots, N\}$, для любого числа $\alpha \in (0, \delta)$ и для любого направления $h \in \mathcal{R}_n$, $\|h\| = 1$, благодаря дифференцируемости функций $J_i(\cdot)$, $i \in \{0, 1, \ldots, N\}$ можно записать соотношение

$$J_i(u_0 + \alpha h) = J_i(u_0) + \alpha (J_i'(u_0), h) + o_i(h, \alpha),$$

где $\frac{o_i(h,\alpha)}{\alpha} \to 0$ при $\alpha \to 0$ равномерно по $h \in \mathcal{R}_n$, $\|h\| = 1$. В силу утверждения 1, найдутся такое число α_0, $0 < \alpha_0 < \delta$, что для всех $0 < \alpha < \alpha_0$ выполнено

$$J(u_0 + \alpha h) = \max_{i \in R(u_0)} \{J_i(u_0) + \alpha (J_i'(u_0), h) + o_i(h, \alpha)\} \le J(u_0) + \alpha \max_{i \in R(u_0)} (J_i'(u_0), h) + \max_{i \in R(u_0)} o_i(h, \alpha) \tag{1}$$

Далее воспользуемся следующим свойством числовых неравенств: для любого набора чисел β_i и γ_i, $i = 0, 1, \ldots, N$,

$$\max_{i\in\{0,1,\ldots,N\}}(\beta_i + \gamma_i) \geq \max_{i\in\{0,1,\ldots,N\}} \beta_i + \max_{i\in R} \gamma_i, \tag{2}$$

где R — подмножество тех индексов i из множества $\{0, 1, \ldots, N\}$, для которых $\beta_i = \max\limits_{k\in\{0,1,\ldots,N\}} \beta_k$: $R = \{i : \beta_i = \max\limits_{k\in\{0,1,\ldots,N\}} \beta_k\}$. Это неравенство выполняется, так как переменная β_i как функция индекса i на множестве R постоянна и равна $c = \max\limits_{k\in\{0,1,\ldots,N\}} \beta_k$, и поэтому

$$\max_{i\in\{0,1,\ldots,N\}}(\beta_i+\gamma_i) \geq \max_{i\in R}(\beta_i+\gamma_i) = \max_{i\in R}(c+\gamma_i) = c+\max_{i\in R}\gamma_i = \max_{i\in\{0,1,\ldots,N\}}\beta_i+\max_{i\in R}\gamma_i.$$

Благодаря неравенству (2) имеем

$$J(u_0+\alpha h) \geq J(u_0)+\max_{i\in R(u_0)}\{\alpha(J_i'(u_0),h)+o_i(h,\alpha)\} \geq J(u_0)+\alpha\max_{i\in R(u_0)}(J_i'(u_0),h)+\max_{i\in R(u_0)}o_i(h,\alpha). \tag{3}$$

Таким образом, из (1) и (3) получим неравенство

$$\min_{i\in R(u_0)} o_i(h,\alpha) \leq J(u_0+\alpha h) - J(u_0) - \alpha\max_{i\in R(u_0)}(J_i'(u_0),h) \leq \max_{i\in R(u_0)} o_i(h,\alpha).$$

Разделив все части этого неравенства на $\alpha > 0$ и устремив $\alpha \to +0$, получим утверждение теоремы, а также соотношение

$$J(u_0+\alpha h) = J(u_0) + \alpha\frac{\partial J(u_0)}{\partial h} + o(h,\alpha),$$

где $\frac{o_i(h,\alpha)}{\alpha} \to 0$ при $\alpha \to 0$ равномерно по $h \in \mathcal{R}_n$, $\|h\| = 1$.

Необходимые условия минимакса. Сформулируем условия, необходимые для достижения минимума функции $J(\cdot)$ для случая, когда все функции $J_i(\cdot)$, $i = 0, 1, \ldots, N$, непрерывны и дифференцируемы на всем пространстве R_n.

Теорема 2. Пусть $J_i(\cdot)$, $i = 0, 1, \ldots, N$, непрерывны и дифференцируемы на $\mathcal{R}_n$ и при некотором $v \in \mathcal{R}_n$ лебегово множество

$$\Lambda_{J(v)} = \{u \in \mathcal{R}_n : \ J(u) \leq J(v)\}$$

ограничено. Тогда для того, чтобы точка $u_* \in \mathcal{R}_n$ была точкой минимума функции $J(\cdot)$ на $\mathcal{R}_n$, необходимо, а в случае выпуклости $J(\cdot)$ на $\mathcal{R}_n$ — и достаточно, чтобы выполнялось неравенство

$$\inf_{h\in\mathcal{R}_n,\|h\|=1}\max_{i\in R(u_*)}(J_i'(u_*),h) \geq 0, \tag{4}$$

или, что то же самое,

$$\inf_{h\in\mathcal{R}_n,\|h\|=1}\frac{\partial J(u_*)}{\partial h} \geq 0. \tag{5}$$

Доказательство. Докажем сначала необходимость. Пусть u_* — точка минимума функции $J(\cdot)$, и предположим, что неравенство (5) не выполняется. Тогда найдется такое направление $h_1 \in \mathcal{R}_n$, $\|h_1\| = 1$, что $\frac{\partial J(u_*)}{\partial h_1} = -a < 0$, и

$$J(u_* + \alpha h_1) = J(u_*) + \alpha \frac{\partial J(u_*)}{\partial h_1} + o(h_1, \alpha).$$

Зафиксируем $\alpha_1 > 0$ такое, что $|o(h_1, \alpha_1)| \leq \frac{a}{2}\alpha_1$, при этом $J(u_* + \alpha_1 h_1) \leq J(u_*) - \frac{a}{2}\alpha_1$, что противоречит предположению о том, что u_* — точка минимума функции $J(\cdot)$. Следовательно, неравенство (5) в точке минимума должно выполняться.

Докажем теперь достаточность условия (4), или эквивалентного ему условия (5). Пусть $J(\cdot)$ — выпукла на $\mathcal{R}_n$ и в некоторой точке $u_* \in \mathcal{R}_n$ выполнено (4). Покажем, что u_* — точка минимума функции $J(\cdot)$ на $\mathcal{R}_n$. Предположим противное, то есть что существует такая точка $u_0 \in \mathcal{R}_n$, в которой $J(u_0) < J(u_*)$. Из выпуклости $J(\cdot)$ на $\mathcal{R}_n$ следует, что

$$J(u_* + \alpha(u_0 - u_*)) \leq (1 - \alpha)J(u_*) + \alpha J(u_0).$$

Выберем $h_1 = \frac{u_0 - u_*}{\|u_0 - u_*\|}$, и найдем $\frac{\partial J(u_*)}{\partial h_1}$. Для $0 < \alpha < \|u_0 - u_*\|$ в силу выпуклости функции $J(\cdot)$ имеем

$$\frac{\partial J(u_*)}{\partial h_1} = \lim_{\alpha \to 0} \frac{1}{\alpha}(J(u_* + \alpha h_1) - J(u_*)) = \lim_{\alpha \to 0} \frac{1}{\alpha}(J(u_* + \frac{\alpha}{\|u_0 - u_*\|}(u_0 - u_*)) - J(u_*)) \leq$$

$$\leq \lim_{\alpha \to +0} \frac{1}{\alpha} \left[\left(1 - \frac{\alpha}{\|u_0 - u_*\|}\right) J(u_*) + \frac{\alpha}{\|u_0 - u_*\|} J(u_0) - J(u_*) \right] =$$

$$= \frac{1}{\|u_0 - u_*\|}(J(u_0) - J(u_*)) < 0,$$

что противоречит (4). Это противоречие и доказывает теорему.

Определение. Точка $u \in \mathcal{R}_n$, в которой выполнены условия (4) или (5), называется стационарной точкой функции $J(\cdot)$.

Дадим геометрическую интерпретацию необходимого условия минимакса. Для этого при фиксированном $u \in \mathcal{R}_n$ образуем множество $H(u)$, состоящее из первых производных функций $J_i(\cdot)$, $i \in R(u)$, то есть из тех векторов $J_i'(u) \in \mathcal{R}_n$, для которых индексы i выбраны так, что на них достигается максимум из всех чисел $J_i(u)$, $i = 0, 1, \ldots, N$:

$$H(u) = \{J_i'(u) : \ i \in R(u)\}.$$

Обозначим $L(u)$ выпуклую оболочку $H(u)$:

$$L(u) = \left\{ v \in \mathcal{R}_n : \ v = \sum_{i \in R(u)} \alpha_i J_i'(u), \ \alpha_i \geq 0, i \in R(u), \sum_{i \in R(u)} \alpha_i = 1 \right\}.$$

Это множество выпукло и замкнуто в $\mathcal{R}_n$.

Теорема 3. Неравенство (4) эквивалентно включению

$$0 \in L(u_*). \tag{6}$$

Доказательство. Покажем, что если выполнено (4), то выполнено и (6). Допустим, что это не так, то есть $0 \notin L(u)$. Тогда по теореме об отделимости существует вектор $h_0 \in \mathcal{R}_n$ единичной нормы и число $a > 0$, такие, что для любого $v \in L(u_*)$ выполняется

$$(v, h_0) \le -a < 0,$$

то есть существует плоскость с вектором нормали h_0, такая, что точка $u = 0$ и выпуклое множество $L(u_*)$ лежат от нее по разные стороны. Так как $H(u_*) \subset L(u_*)$, то последнее неравенство выполнено для любого вектора $J_i'(u_*)$, $i \in R(u)$, что противоречит условию (4).

Покажем теперь, что из (6) следует (4). Предположим противное, тогда существует вектор $h_0 \in \mathcal{R}_n$ единичной нормы, такой, что для всех $i \in R(u_*)$

$$(J_i'(u_*), h_0) \le -a < 0,$$

и для произвольного вектора $v \in L(u_*)$ в силу определения множества $L(u_*)$ выполнено

$$(v, h_0) = \left(\sum_{i \in R(u_*)} \alpha_i J_i'(u_*), h_0 \right) = \sum_{i \in R(u_*)} \alpha_i (J_i'(u_*), h_0) < -a \sum_{i \in R(u_*)} \alpha_i = -a < 0.$$

Отсюда следует, что вектор $v = 0$ не может принадлежать множеству $L(u_*)$, что противоречит (6).

Пример. Пусть требуется найти минимум функции $J(\cdot)$, определенной для каждого $u \in \mathcal{R}_n$ соотношением $J(u) = \max\{\|A(u - u_0)\|^2, \|Bu\|^2\}$, где $A \in (R_n \to \mathcal{R}_m)$ и $B \in (R_n \to \mathcal{R}_k)$ — невырожденные линейные операторы.

Воспользовавшись необходимыми условиями минимума, найдем, что в точке минимума должно выполняться одно из соотношений:

$$\|A(u - u_o)\|^2 > \|Bu\|^2 \quad \text{и} \quad A^*A(u - u_0) = 0, \tag{7}$$

$$\|A(u - u_o)\|^2 < \|Bu\|^2 \quad \text{и} \quad B^*Bu = 0 \tag{8}$$

или

$$\|A(u - u_o)\|^2 = \|Bu\|^2 \quad \text{и} \quad \alpha A^*A(u - u_0) + (1 - \alpha)B^*Bu = 0 \tag{9}$$

для некоторого $\alpha \in [0,])$.

Но соотношения (a) и (b) невозможны в силу неотрицательности нормы, а из соотношения (c) получаем

$$u = u(\gamma) = (A^*A + \gamma B^*B)^{-1} A^*Ax_0,$$

где $\gamma = \frac{1-\alpha}{\alpha} \in (0, \infty)$ — корень уравнения

$$\|A(u(\gamma) - u_o)\|^2 = \|Bu(\gamma)\|^2.$$

Заметим, что параметр α при $u_0 \neq 0$ не может равняться ни нулю, ни единице, так как при этом из (9) следует равенство $\|A(u - u_o)\|^2 = \|Bu\|^2 = 0$, эквивалентное равенству $A(u - u_o) = Bu = 0$, не выполняющемуся в силу невырожденности A и B.

Численные методы решения задачи на минимакс. Рассмотрим методы, позволяющие найти минимум функции $J(\cdot)$ на $\mathcal{R}_n$. Для этого определим ряд понятий.

Определение. Вектор $h(u) \in \mathcal{R}_n$, $\|h(u)\| = 1$, называется направлением наискорейшего спуска функции максимума $J(\cdot)$ в точке $u \in \mathcal{R}_n$, если

$$\frac{\partial J(u)}{\partial h(u)} = \min_{h \in \mathcal{R}_n, \|h\|=1} \frac{\partial J(u)}{\partial h}.$$

Обозначим

$$\psi(u) = \min_{h \in \mathcal{R}_n, \|h\|=1} \max_{i \in R(u)} (J_i'(u), h), \quad u \in \mathcal{R}_n.$$

Утверждение 2. Если в некоторой точке $u \in \mathcal{R}_n$ выполнено

$$\psi(u) < 0,$$

то функция $J(\cdot)$ имеет в точке $h \in \mathcal{R}_n$ единственное направление наискорейшего спуска.

На этом факте основан метод наискорейшего спуска предназначенный для поиска стационарных точек функции $J(\cdot)$, к описанию которого мы сейчас приступаем. Напомним, что если функция $J(\cdot)$ выпукла на $\mathcal{R}_n$, то в стационарной точке достигается точная нижняя грань функции максимума $J(\cdot)$.

Пусть лебегово множество $\Lambda_{J(v)}$ ограничено при некотором $v \in \mathcal{R}_n$, тогда в силу непрерывности $J(\cdot)$ оно является и замкнутым множеством. Возьмем точку $u_0 \in \Lambda_{J(v)}$ в качестве точки начального приближения. Если $\psi(u_0) \geq 0$, то u_0 — стационарная точка функции $J(\cdot)$, и процедура ее поиска на этом закончена. В противном случае, в силу утверждения 2, существует единственное направление наискорейшего спуска, которое обозначим $h(u_0)$, $\|h(u_0)\| = 1$; для него построим луч

$$u = u_0 + \alpha h(u_0), \quad \alpha \geq 0,$$

и найдем такое число α_0, при котором выполнено равенство

$$J(u_0 + \alpha_0 h(u_0)) = \min_{\alpha \geq 0} J(u_0 + \alpha h(u_0)).$$

Минимум здесь достигается в силу непрерывности $J(\cdot)$ и ограниченности и замкнутости лебегова множества $\Lambda_{J(v)}$. После этого точка $u_1 = u_0 + \alpha h(u_0)$ становится новой точкой начального приближения, и вся процедура повторяется с заменой u_0 на u_1.

Очевидно, $\{u_k\}$ является минимизирующей последовательностью для $J(\cdot)$. Если в результате мы пришли к тому, что на некотором шаге $k_0 < \infty$ выполнено неравенство $\psi(u_{k_0}) \geq 0$, то u_{k_0} является искомой стационарной точкой. Однако, возможны ситуации, когда минимизирующая последовательность $\{u_k\}$ не содержит предельных точек, являющихся стационарными точками функции $J(\cdot)$.

Пример. Пусть $u = (u_1, u_2) \in \mathcal{R}_2$, заданы функции $J_1(u_1, u_2) = -5u_1 + u_2$, $J_2(u_1, u_2) = 4u_2 + u_1^2 + u_2^2$, $J_3(u_1, u_2) = 5u_1 + u_2$, и $J(u_1, u_2) = \max\limits_{i=1,2,3} J_i(u_1, u_2)$, $-\infty < u_1 < \infty$, $-\infty < u_2 < \infty$. Функции $J_i(\cdot)$ выпуклы на $\mathcal{R}_2$, поэтому и функция максимума $J(\cdot)$ также выпукла на $\mathcal{R}_2$, причем

$$J(u) = \begin{cases} J_1(u), & \text{если} \quad u \in A_1 \\ J_2(u), & \text{если} \quad u \in A_2 \\ J_3(u), & \text{если} \quad u \in A_3 \end{cases}, \quad u \in \mathcal{R}_2,$$

где области $A_i \subset \mathcal{R}_2$, $i = 1, 2, 3$, изображены на рис.1.

Точка $\overline{u}_0 = (0, 0) \in \mathcal{R}_2$, как легко проверить, воспользовавшись теоремой 3, не является стационарной точкой для функции $J(\cdot)$, а точка $\overline{u}_1 = (0, -3) \in \mathcal{R}_2$ — стационарна. Если выбрать начальное приближение так, как показано на рис.2, то минимизирующая последовательность состоит из точек, изображенных на рис.2; она сходится к точке $\overline{u}_0 \in \mathcal{R}_2$, не являющейся стационарной, причем сама точка $\overline{u}_0$ не принадлежит этой последовательности, что не позволяет минимизирующей последовательности двигаться вдоль направления наискорейшего спуска из точки $\overline{u}_0$ вдоль хорды, соединяющей точки $\overline{u}_0$ и $\overline{u}_1$. Избавиться от недостатков метода наискорейшего спуска можно, расширив множество функций $J_i(\cdot)$, определяющих направление наискорейшего спуска в заданной точке $u \in \mathcal{R}_n$. Действительно, в рассмотренном примере в точках, близких к точке $\overline{u}_0 = (0, 0)$, значения функций $J_i(u)$, $i = 1, 2, 3$, мало отличаются между собой, и если бы направление наискорейшего спуска $h(u)$ в точках из окрестности $O_\delta(\overline{u}_0) \subset \mathcal{R}_2$ вычислялось бы не только как направления градиентов функций из множества $J_i(u)$, $i = R(u)$, на которых достигается наибольшая скорость убывания функции $J(\cdot)$ в точке u, а учитывались и скорости убывания функций, значения которых $J_l(u)$, $l \notin R(u)$,

близки к значению $J(\cdot)$ в точке u, то направление спуска от точки $\overline{u}_0$ к точке $\overline{u}_1$ "ощущалось"бы и в окрестности точки $\overline{u}_0$.

Пусть, как и раньше, $J_i(\cdot)$, $i = 0, 1, \ldots, N$, непрерывно дифференцируемы на $\mathcal{R}_n$. Для заданного $\varepsilon \geq 0$ и фиксированной точки $u \in \mathcal{R}_n$ обозначим

$$R_\varepsilon(u) = \{i | J(u) - J_i(u) \leq \varepsilon\}.$$

Ясно, что $R_0(u) = R(u)$, и $R_{\varepsilon'}(u) \supseteq R_\varepsilon(u)$, если $\varepsilon' \geq \varepsilon$. Введем функцию

$$\psi_\varepsilon(u) = \min_{h \in \mathcal{R}_n, \|h\|=1} \max_{i \in R_\varepsilon(u)} (J_i'(u), h), \quad u \in \mathcal{R}_n.$$

Точно так же, как и для $R_\varepsilon(u)$, $\psi_0(u) = \psi(u)$, и при фиксированном $u \in \mathcal{R}_n$ функция $\psi_\varepsilon(u)$ как функция ε является неубывающей и кусочно постоянной.

Определение. Точку $u_* \in \mathcal{R}_n$ назовем ε-стационарной для функции $J(\cdot)$, если для нее выполнено равенство

$$\psi_\varepsilon(u_*) \geq 0.$$

Вектор $h_\varepsilon(u) \in \mathcal{R}_n$, $\|h_\varepsilon(u)\| = 1$, назовем направлением ε-наискорейшего спуска функции $J(\cdot)$ в точке $u \in \mathcal{R}_n$, если

$$\max_{i \in R_\varepsilon(u)} (J'(u), h_\varepsilon(u)) = \psi_\varepsilon(u).$$

Точно так же, как при геометрической интерпретации необходимых условий стационарности точки $u_* \in \mathcal{R}_n$, положим

$$H_\varepsilon(u) = \{u = J_i'(u),\ i \in R_\varepsilon(u)\},$$

и обозначим $L_\varepsilon(u) = \operatorname{co} H_\varepsilon(u)$ выпуклую оболочку, натянутую на множество $H_\varepsilon(u)$.

Справедливы следующие утверждения.

Утверждение 3. Для того, чтобы $u_* \in \mathcal{R}_n$ была ε-стационарной точкой функции $J(\cdot)$, необходимо и достаточно, чтобы нулевой элемент $u_0 = 0 \in \mathcal{R}_n$ принадлежал множеству $L_\varepsilon(u)$.

Утверждение 4. Если точка $u \in \mathcal{R}_2$ не является ε-стационарной точкой функции $J(\cdot)$, то есть если $\psi_\varepsilon(u) < 0$, то функция $J(\cdot)$ в точке $u \in \mathcal{R}_n$ имеет единственное направление ε-наискорейшего спуска $h_\varepsilon(u) \in \mathcal{R}_n$.

Опишем метод последовательных приближений для поиска ε-стационарных точек функции $J(\cdot)$ на пространстве $\mathcal{R}_n$, и укажем, как можно его использовать для нахождения стационарных точек. Этот метод строится так же, как и метод наискорейшего спуска, только вместо направления наискорейшего спуска $h(u_k)$ на каждом шаге выбирается направление ε-наискорейшего спуска $h_\varepsilon(u_k)$.

Так же, как и ранее, будем считать, что при некотором $v \in \mathcal{R}_n$ лебегово множество $\Lambda_{J(v)}$ функции $J(\cdot)$ ограничено и замкнуто. Возьмем точку $u_0 \in \Lambda_{J(v)}$ в качестве точки начального приближения. Если $\psi_\varepsilon(u_0) \geq 0$, то u_0 —ε-стационарная

точка функции $J(\cdot)$, и процедура ее поиска на этом закончена. В противном случае, в силу утверждения 4, существует единственное направление ε-наискорейшего спуска, которое обозначим $h_\varepsilon(u_0)$, $\|h_\varepsilon(u_0)\| = 1$; для него построим луч

$$u = u_0 + \alpha h_\varepsilon(u_0), \quad \alpha \geq 0,$$

и найдем такое число $\alpha_0 > 0$, при котором выполнено равенство

$$J(u_0 + \alpha_0 h_\varepsilon(u_0)) = \min_{\alpha \geq 0} J(u_0 + \alpha h_\varepsilon(u_0)).$$

Минимум здесь достигается в силу непрерывности $J(\cdot)$ и ограниченности и замкнутости лебегова множества $\Lambda_{J(v)}$. После этого точка $u_1 = u_0 + \alpha h_\varepsilon(u_0)$ становится новой точкой начального приближения, и вся процедура повторяется с заменой u_0 на u_1.

Справедливо следующее утверждение.

Утверждение 5. Любая предельная точка последовательности $\{u_k\} \subset \mathcal{R}_n$, полученной методом последовательных приближений, является ε-стационарной точкой функции $J(\cdot)$ на $\mathcal{R}_n$.

Алгоритм метода последовательных приближений может быть использован для нахождения стационарной точкой функции $J(\cdot)$ на $\mathcal{R}_n$. Для этого укажем точку $u_{0,0} \in \mathcal{R}_n$ начального приближения и зафиксируем числа $\varepsilon_0 > 0$ и $\rho_0 > 0$, после чего методом последовательных приближений за конечное число шагов l_0 построим точку $u_{0,l_0} \in \mathcal{R}_n$, для которой выполняется неравенство

$$\psi_{\varepsilon_0}(u_{0,l_0}) \geq -\rho_0.$$

Затем положим $\varepsilon_1 = \varepsilon_0/2$, $\rho_1 = \rho_0/2$, и зададим новую точку начального приближения $u_{1,0} = u_{0,l_0} \in \mathcal{R}_n$, для которой повторим процедуру, в результате которой за l_1 число шагов получим точку $u_{1,l_1} \in \mathcal{R}_n$, для которой

$$\psi_{\varepsilon_1}(u_{1,l_1}) \geq -\rho_1.$$

Утверждение 6. Любая предельная точка последовательности $\{u_{k,l_k}\} \subset \mathcal{R}_n$ является стационарной точкой функции $J(\cdot)$ на $\mathcal{R}_n$.

Глава 7. Общая непрерывная задача на минимакс.

Необходимые условия минимакса. Описанные в предыдущей главе результаты обобщаются на случай, когда индекс i меняется не в конечном множестве, а является элементом подмножества m-мерного евклидова пространства, и функция максимума $J(\cdot)$ минимизируется не на всем пространстве, а на некотором его подмножестве. Опишем кратко эти результаты.

Пусть $\mathcal{U}' \subset \mathcal{R}_n$ — открытое, а $\mathcal{V} \subset \mathcal{R}_n$ — ограниченное и замкнутое подмножества евклидовых пространств $\mathcal{R}_n$ и $\mathcal{R}_m$ размерности $n < \infty$ и $m < \infty$ соответственно, $J(u, v)$, $u \in \mathcal{U}'$, $v \in \mathcal{V}$ — функция, непрерывно дифференцируемая по $u \in \mathcal{U}'$. Рассмотрим задачу минимизации функции

$$J(u) = \max_{v \in \mathcal{V}} J(u, v), \qquad u \in \mathcal{U}',$$

на ограниченном замкнутом подмножестве $\mathcal{U} \subset \mathcal{U}'$.

Далее полагаем, что функция $J(u, v)$, $u \in \mathcal{U}'$, $v \in \mathcal{V}$, непрерывна вместе со своими производными по u по совокупности переменных на множестве $\mathcal{U}' \otimes \mathcal{V} \subset \mathcal{R}_n \otimes \mathcal{R}_m$.

Свойства функции максимума $J(\cdot)$.

1. Функция $J(\cdot)$ непрерывна на множестве $\mathcal{U}'$.

2. Если при каждом $v \in \mathcal{V}$ функция $J(u, v)$ выпукла по u на выпуклом множестве $\mathcal{U} \subset \mathcal{U}'$, то и функция $J(\cdot)$ выпукла на $\mathcal{U}$.

3. Если $\mathcal{U} \subset \mathcal{U}'$ — замкнуто, и при некотором $u_0 \in \mathcal{U}$ лебегово множество $\Lambda_{J(u_0)}$ функции $J(\cdot)$ ограничено, то $J(\cdot)$ достигает своей точной нижней грани на множестве $\mathcal{U}$.

Утверждение 1. Функция $J(\cdot)$ имеет в каждой точке $u \in \mathcal{U}'$ производную по любому направлению $h \in \mathcal{R}_n$, $\|h\| = 1$, причем

$$\frac{\partial J(u)}{\partial h} = \max_{v \in R(u)} (J'_u(u, v), h),$$

где $J'_u(u, v)$ — производная функции $J(u, v)$, взятая по переменной $u \in \mathcal{U}'$ при фиксированном $v \in \mathcal{V}$, а

$$R(u) = \{v \in \mathcal{V} : \ J(u, v) = J(u)\}.$$

С учетом того, что минимум функции $J(\cdot)$ ищется не на всем пространстве $\mathcal{R}_n$, а лишь на его ограниченном замкнутом подмножестве $\mathcal{U}$, необходимые условия минимума сформулируем следующим образом.

Утверждение 2. Для того, чтобы функция $J(\cdot)$ достигала максимального значения на выпуклом замкнутом множестве $\mathcal{U} \subset \mathcal{U}'$ в точке $u_* \in \mathcal{U}$, необходимо, а в случае выпуклости функции $J(\cdot)$ на $\mathcal{U}$ — и достаточно, чтобы выполнялось равенство

$$\inf_{w \in \mathcal{U}} \max_{v \in R(u_*)} (J'_u(u_*, v), w - u_*) = 0, \tag{1}$$

причем если $\mathcal{U} \equiv \mathcal{R}_n$, то условие (1) эквивалентно условию

$$\min_{h \in \mathcal{R}_n \ \|h\|=1} \max_{v \in R(u_*)} (J'_u(u_*, v), h) \geq 0.$$

Последнее условие почти дословно повторяет необходимое условие минимума для дискретной минимаксной задачи, а условие (1) обобщает его, учитывая, что точка u_* может находиться на границе области $\mathcal{U}$.

Минимакс и максимин. Сформулируем условия, при которых в задаче на минимакс можно "поменять местами"условия минимизации и максимизации, так, что выполняется равенство

$$\min_{u\in\mathcal{U}}\max_{v\in\mathcal{V}} J(u,v) = \max_{v\in\mathcal{V}}\min_{u\in\mathcal{U}} J(u,v). \tag{2}$$

Определение. Точка $(u_*, v_*) \in \mathcal{U}\otimes\mathcal{V}$ называется седловой точкой функции $J(u,v)$ на множестве $\mathcal{U}\otimes\mathcal{V}$, если выполнены неравенства

$$J(u_*, v) \le J(u_*, v_*) \le J(u, v_*).$$

Утверждение 3. Существование седловой точки функции $J(u,v)$ на множестве $\mathcal{U}\otimes\mathcal{V}$ эквивалентно выполнению равенства (2).

Укажем конструктивные достаточные условия существования седловой точки функции $J(u,v)$ на множестве $\mathcal{U}\otimes\mathcal{V}$, или, что то же самое, выполнения равенства (2).

Утверждение 4. Пусть функция $J(\cdot,\cdot)$ непрерывна вместе со своими производными $J'_u(\cdot,\cdot)$ на множестве $\mathcal{U}'\otimes\mathcal{V}$, где $\mathcal{U}'\subset\mathcal{R}_n$ — открытое, а $\mathcal{V}\subset\mathcal{R}_m$ — ограниченное и замкнутое выпуклые множества. Если при каждом фиксированном $u_0\in\mathcal{U}'$ функция $J(u_0,v)$ вогнута по $v\in\mathcal{V}$, и при каждом фиксированном $v_0\in\mathcal{V}$ функция $J(u,v_0)$ выпукла по $u\in\mathcal{U}'$, то выполняются равенства (2).

Минимакс и метод минимальных модулей. Часто при решении уравнений вида

$$y = Ax, \tag{3}$$

где x и y — векторы линейных пространств размерности N и n соответственно, а A — линейный оператор, вместо решения уравнения проще найти такой вектор x, при котором Ax как можно ближе к y в метрике пространства C с нормой $\|y\|_C = \max\limits_{j=1,\dots,n}|y_i|$.

Пусть линейный оператор задан матрицей A, строки которой обозначим a_i, $i=1,\dots,n$; каждая строка имеет N координат. Тогда уравнение (3) примет вид

$$y_i = (a_i, x), \quad i=1,\dots,n.$$

Вместо решения системы уравнений (3) рассмотрим задачу на минимакс

$$\min_x \max_{i=1,\dots,n} |y_i - (a_i, x)|. \tag{4}$$

Эта задача на минимакс сводится к задаче линейного программирования.

Действительно, для вычисления модуля $|d|$ требуется найти минимальное из чисел z, для которых выполнены неравенсвта $z\ge d$ и $z\ge -d$.

Далее, для вычисления $z = \max\limits_{i=1,\dots,n}|y_i-(a_i,x)|$ при фиксированном x требуется найти минимальное из чисел z, для которых выполнены неравенства $z\ge y_i-(a_i,x)$ и $z\ge -y_i+(a_i,x)$, $i=1,\dots,n$.

Если x — любой вектор из N-мерного линейного пространства, и требуется найти минимум по x в (4), то при выборе минимального числа z, для которого выполнены указанные выше неравенства, следует разрешить изменяться и самим векторам x и искать минимальное из чисел z при переменных $(z, x_1, \ldots, x_N)$. Это можно опеспечить, введя два вектора, s и t, с координатами $s = (z, x_1, \ldots, x_N)$, $t = (1, 0, \ldots, 0)$. Скалярное произведение (s, t) равно z, и минимизация скалярного произведения (s, t) выбором вектора $s = (z, x_1, \ldots, x_N)$ при ограничениях $z \geq y_i - (a_i, x)$ и $z \geq -y_i + (a_i, x)$, $i = 1, \ldots, n$ приведет к тому, что первая координата z решения s задачи линейного программирования даст значение минимума (4), а последующие N координат — координаты вектора x, на котором этот минимум достигается.

Часть 4. Минимизация функций на линейных нормированных пространствах.

В этом разделе рассматриваются задачи на минимум функций, заданных на пространстве $\mathcal{C}$ непрерывных функций, а также на пространстве $\mathcal{C}^1$ функций, непрерывных со своей первой производной.

Будем рассматривать пространство $\mathcal{C}$ как линейное пространство, состоящее из функций, заданных и непрерывных на отрезке $[a,b] \in \mathcal{R}_1$. Норма элемента u пространства $\mathcal{C}$ определяется как максимальное значение модуля функции $u(t)$ на отрезке $[a,b]$:

$$\|u\| = \max_{t\in[a,b]} |u(t)|.$$

Пространство $\mathcal{C}^1$ состоит из функций, заданных на отрезке $[a,b] \in \mathcal{R}_1$ и имеющих на $[a,b]$ непрерывную первую производную. Норма элемента u пространства $\mathcal{C}^1$ определяется соотношением

$$\|u\|_1 = \max_{t\in[a,b]} |u(t)| + \max_{t\in[a,b]} |u'(t)|.$$

Далее пространство, на котором задана минимизируемая функция $J(\cdot)$, будем обозначать символом $\mathcal{R}$, имея в виду, что в зависимости от рассматриваемой ситуации либо $\mathcal{R} = \mathcal{C}$, либо $\mathcal{R} = \mathcal{C}^1$. Функцию $J(\cdot)$, заданную на $\mathcal{R}$ или на его подмножестве $\mathcal{U} \in \mathcal{R}$, будем называть функционалом, чтобы отличать от ее аргумента: функция $J(\cdot)$ определена на элементах $u \in \mathcal{R}$ — функциях $u(\cdot)$, заданных на отрезке $[a,b]$. Непрерывность функции $J(\cdot)$ на $\mathcal{R}$ определяется точно так же, как для функций, заданных на евклидовых пространствах, то есть $J(\cdot)$ непрерывна в точке $u \in \mathcal{R}$, если для любой последовательности аргументов $\{u_k\} \in \mathcal{R}$, сходящейся к $u \in \mathcal{R}$ по норме пространства $\mathcal{R}$, соответствующая числовая последовательность значений функционала $J(u_k)$ сходится к $J(u)$. Заметим, что непрерывный в точке $u \in \mathcal{C}^1$ функционал, определенный на пространстве $\mathcal{C}^1$, может терпеть разрыв в этой же точке, если он рассматривается на подмножестве непрерывно дифференцируемых функций пространства $\mathcal{C}$, так как последовательность $\{u_k\}$, сходящаяся по норме пространства $\mathcal{C}$, может не иметь предела по норме пространства $\mathcal{C}^1$.

Подмножество $\mathcal{U} \subset \mathcal{R}$ назовем линейным множеством пространства $\mathcal{R}$, если из $u_1, u_2 \in \mathcal{U}$ следует, что $\alpha u_1 + \beta u_2 \in \mathcal{U}$ для любых чисел α и β.

Функционал $\varphi(\cdot)$, определенный на линейном множестве пространства $\mathcal{R}$, называется линейным, если он

1. непрерывен;
2. $\varphi(\alpha u_1 + \beta u_2) = \alpha\varphi(u_1) + \beta\varphi(u_2)$ для любых $u_1, u_2 \in \mathcal{U}$ и для любых чисел α и β.

Приведем пример функционала, для которого выполнено второе из этих условий, но не выполнено первое. Пусть на множестве $\mathcal{U}$ непрерывно дифференцируемых функций, определенных на отрезке $[0, \pi]$, задан функционал

$$J(u) = \left.\frac{du}{dt}\right|_{t=\frac{\pi}{2}}.$$

Его область определения $\mathcal{U}$ — линейное множество в $C[0, \pi]$, так как произведение дифференцируемой функции на число и сумма дифференцируемых функций дифференцируемы. На области определения $\mathcal{U}$ в силу линейности интеграла выполнены условия 2 приведенного выше определения. Однако данный функционал не является непрерывным в нуле. Действительно, рассмотрим последовательности функций $\{u_k(\cdot)\}$, $u_k(t) = \frac{\cos(kt)}{k}$, $t \in [0, \pi]$, $k = 1, 2, \ldots, \infty$. Эта последовательность при $k \to \infty$ сходится в $C[0, \pi]$ к нулю, но последовательность

$$J(u_k) = -\sin\frac{k\pi}{2}$$

не имеет предела при $k \to \infty$.

Рассмотрим приращение $\Delta J(h)$ функционала $J(\cdot)$ в точке $u \in \mathcal{R}$:

$$\Delta J(h) = J(u + h) - J(u), \qquad h \in \mathcal{R}.$$

если

$$\Delta J(h) = \varphi(h) + o(u, \|h\|),$$

где $\varphi(\cdot)$ — линейный функционал аргумента $h \in \mathcal{R}$, а $\frac{o(u,\|h\|)}{\|h\|} \to 0$ при $h \to 0$, то функционал $J(\cdot)$ в точке $u \in \mathcal{R}$ называется дифференцируемым, а функционал $\varphi(\cdot)$ называется его вариацией, или дифференциалом.

Далее будем рассматривать функционал $J(\cdot)$, определенный на некотором множестве дифференцируемых функций из $\mathcal{C}^1$. Сами эти функции можно считать и элементами пространства $\mathcal{C}$, и элементами пространства $\mathcal{C}^1$. В соответствии с этим можно рассмотреть две возможности. Будем говорить, что $J(\cdot)$ достигает в точке $u = u_*$ слабого минимума, если найдется такое число $\varepsilon > 0$, такое, что для всех элементов $u \in \mathcal{C}^1$, для которых $J(\cdot)$ определен и таких, что $\|u - u_*\|_1 \leq \varepsilon$, выполняется неравенство

$$J(u) - J(u_*) \geq 0. \qquad (1^*)$$

Соответственно, $J(\cdot)$ достигает в точке $u = u_*$ сильного минимума, если неравенство (1^*) выполнено для всех элементов u, принадлежащих области определения функционала $J(\cdot)$ и таких, что $\|u - u_*\| \leq \varepsilon$.

Утверждение 1. Любой сильный минимум является в то же время и слабым минимумом.

Действительно, если $\|u-u_*\|_1 \le \varepsilon$, то и подавно $\|u-u_*\| \le \varepsilon$, поэтому если $J(u_*)$ есть минимальное значение функционала $J(\cdot)$ по отношению к тем u, для которых $\|u-u_*\| \le \varepsilon$, то $J(u_*)$ тем более является минимальным значениям по отношению к тем u, для которых $\|u-u_*\|_1 \le \varepsilon$.

Глава 8. Вариационное исчисление.

Постановка задачи. Уравнение Эйлера. В этой главе мы остановимся на поиске минимума функционалов вида

$$J(u) = \int_a^b F(t,u,u')dt, \tag{1}$$

где $F(\cdot,\cdot,\cdot)$ — функция, имеющая непрерывные частные производные до второго порядка включительно по всем переменным, а $u(\cdot)$ — функции, заданные на $[a,b]$, имеющие первую производную и удовлетворяющие граничным условиям $u(a)=A$, $u(b)=B$ Класс функций, обладающих указанными свойствами, обозначим $\mathcal{U}$.

Утверждение 2. Для того, чтобы функционал (1) достигал на множестве $\mathcal{U}$ своего минимального значения на элементе $u_* \in \mathcal{U}$, необходимо, чтобы функция $u_*(\cdot)$ удовлетворяла условиям Эйлера

$$F'_u - \frac{d}{dt}F'_{u'} = 0. \tag{2}$$

Интегральные кривые уравнения Эйлера называются экстремалями функционала (1).

Пример 1. Если $u(t)$ — координата материальной точки, $F(t,u,u')$ — функция Лагранжа, то $J(\cdot)$ — действие, а уравнение (1) — уравнение движения, то есть второй закон Ньютона.

Заметим, что в (1) функция $u(\cdot)$ может рассматриваться как функция, заданная на отрезке $[a,b]$ и принимающая значения в евклидовом пространстве $\mathcal{R}_n$, при этом каждая ее координата рассматривается как дифференцируемая функция. В граничных условиях при этом $u(a)=A\in\mathcal{R}_n$, $u(b)=B\in\mathcal{R}_n$.

Пример 2. Пусть кривая в n-мерном евклидовом пространстве задана параметрически: $x_i = u_i(t)$, $i=1,\dots,n$, $t\in[t_1,t_2]$; тогда ее длина вычисляется по формуле

$$l = \int_{t_1}^{t_2} \|\dot u(t)\|dt.$$

В этом примере $F(t,u,u') = \|\dot u\|$.

Для функционалов, заданных на $\mathcal{C}$ или на $\mathcal{C}^1$, наряду с понятием дифференциала вводится понятие вариационной производной в точке $t=t_0\in[a,b]$.

Пусть задан функционал $J(\cdot)$ на множестве $\mathcal{U}$ функционального пространства $\mathcal{R}$, $u = u(t)$, $t \in [a, b]$, — непрерывная функция. Зададим функции u приращение h, отличное от нуля лишь в окрестности точки t_0, и вычислим соответствующее приращение ΔJ функционала $J(\cdot)$: $\Delta J(h, u) = J(u + h) - J(u)$. Обозначим $\Delta s = \int\limits_a^b h(t)dt$ — площадь, ограниченную кривой $h(t)$ и осью t, $t \in [a, b]$, и рассмотрим отношение

$$\frac{J(u + h) - J(u)}{\Delta s}. \tag{3}$$

Пусть теперь $\Delta s \to 0$ так, что и $\|h\| = \max\limits_{a \le t \le b} |h(t)|$, и длина того интервала, на котором $h(\cdot)$ отлична от нуля, стремятся к нулю. Если при этом отношение (3) стремится к некоторому пределу, то этот предел называется вариационной производной функционала $J(\cdot)$ в точке $t = t_0$.

Утверждение 3. Пусть на плоскости вместо переменных (t, u) введены некоторые новые переменные (τ, v) по формулам

$$\tau = \tau(t, u),$$

$$v = v(t, u),$$

причем якобиан этого преобразования отличен от нуля, так, что кривой $u = u(t)$, $t \in [a, b]$, соответствует кривая $v = v(\tau)$, $\tau \in [\alpha, \beta]$. Если $J_1(v) = J(u)$, то для функционала $J(\cdot)$ уравнения Эйлера сохраняют свой вид:

$$\frac{\partial F_1}{\partial v} - \frac{d}{d\tau}\frac{\partial F_1}{\partial v'} = 0;$$

здесь

$$F_1(\tau, v, v') = F\left(t(\tau, v), u(\tau, v), \frac{u_\tau + u_v v'}{t_\tau + t_v v'}\right)(t_\tau + t_v v')$$

— функция, полученная из подынтегральной функции $F(\cdot, \cdot, \cdot)$ в $J(\cdot)$ путем замены переменных под знаком интеграла.

Это утверждение позволяет при минимизации функционала $J(\cdot)$ пользоваться заменой переменных не в уравнениях Эйлера, а прямо в интеграле, представляющим рассматриваемый функционал, а затем уже для нового интеграла записывая уравнения Эйлера.

Условный экстремум. Рассмотрим теперь методы минимизации функционалов при наличии дополнительных ограничений на аргумент $u(\cdot)$. Пусть среди всех кривых $u(t)$, $t \in [a, b]$, удовлетворяющих условиям $u(a) = A$, $u(b) = B$, на которых заданный функционал

$$K(u) = \int\limits_a^b G(t, u, u')dt \tag{4}$$

принимает заданное значение l, требуется найти такую, для которой другой функционал

$$J(u) = \int_a^b F(t, u, u')dt$$

достигает минимума.

Утверждение 4. Если на кривой $u = u(t)$, $a \le t \le b$, достигается минимум функционала $J(\cdot)$, и $K(u) = l$, причем $u(a) = A$, $u(b) = B$, и u не является экстремалью функционала (4), то существует такое число $\lambda = const$, что u является экстремалью функционала

$$\int_a^b (F(t, u, u') + \lambda G(t, u, u'))dt.$$

Канонический вид уравнения Эйлера. Уравнение Эйлера

$$F'_u - \frac{d}{dt}F'_{u'} = 0, \quad F'_u \in \mathcal{R}_n, \quad F'_{u'} \in \mathcal{R}_n,$$

записанное в некотором ортонормированном базисе пространства $\mathcal{R}_n$, является системой n уравнений второго порядка. Ее можно свести к $2n$ обыкновенным дифференциальным уравнениям первого порядка, например, таким образом:

$$\begin{cases} F'_u - \frac{d}{dt}F'_g = 0, \\ \frac{du}{dt} = g \end{cases},$$

где $u \in \mathcal{R}_n$, $g \in \mathcal{R}_n$ — неизвестные функции, а t — независимая переменная. Однако удобен следующий способ сведения к системе $2n$ уравнений, называемой канонической. Обозначим

$$p = F'_{u'} \in \mathcal{R}_n, \tag{5}$$

$$H = -F + (u', p), \tag{6}$$

где $(\cdot, \dot{)}$ — скалярное произведение в евклидовом пространстве $\mathcal{R}_n$. Выразив из (5) u' через t, u и p, примем величины t, u и p за новые переменные вместо прежних t, u и u' и сделаем эту замену в уравнениях Эйлера. Одновременно функцию F, входящую в уравнение Эйлера, выразим через функцию H, связанную с F уравнением (6), в котором u' есть функция аргументов t, u и p. Определенная таким образом функция $H(t, u, p)$ называется функцией Гамильтона, отвечающая функционалу $J(\cdot)$ вида (1). Переменные t, u, p и H, связанные со старыми переменными t, u, u' и F соотношениями (5)-(6), называются каноническими переменными; уравнение Эйлера для них имеет вид

$$\frac{dp}{dt} = -\frac{\partial H}{du}, \qquad \frac{du}{dt} = \frac{\partial H}{dp}. \tag{7}$$

Уравнение Гамильтона-Якоби. Рассмотрим функционал

$$J(u) = \int_a^b F(t, u, u')dt,$$

определенный на кривых, лежащих в области $G \subset \mathcal{R}_n$, и предположим, что через любые точки $A \in G$ и $B \in G$ проходит одна и только одна экстремаль функционала $J(\cdot)$. Величину

$$S = \int_a^b F(t, u, u')dt,$$

где интеграл берется вдоль экстремали, соединяющей точки $A(t_0) = u(t_0))$ и $B(t_1) = u(t_1))$, где $t_0 = a$, $t_1 = b$, называется геодезическим расстоянием между точками A и B. Функция $S = S(A, B)$ определена однозначно координатами точек A и B. Зафиксируем A и рассмотрим $S(t, u)$ как функцию конечной точки $B(t_1) = u(t_1)$. Повторяя рассуждения, проведенные в курсе теоретической механики, прийдем к уравнению для S:

$$\frac{\partial S}{\partial t} + H(t, u, \frac{\partial S}{\partial u}) = 0, \tag{8}$$

называемым уравнением Гамильтона-Якоби. Для уравнения (8) канонические уравнения (7) образуют характеристическую систему.

Утверждение 5. Пусть $S = S(t, u, \alpha)$, $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathcal{R}_m$ — искомое решение уравнения Гамильтона-Якоби, зависящее от α как от параметра. Тогда каждая из производных $\frac{\partial S}{\partial \alpha_i}$, $i = 1, \ldots, m$, является первым интегралом системы уравнений Эйлера

$$\frac{dp}{dt} = -\frac{\partial H}{du}, \qquad \frac{du}{dt} = \frac{\partial H}{dp},$$

то есть $\frac{\partial S}{\partial \alpha_i} = const$, $i = 1, \ldots, m$, вдоль каждой экстремали.

Это утверждение означает, что если нам известен полный интеграл уравнения Гамильтона-Якоби, то есть его решение $S = S(t, u, \alpha)$, $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathcal{R}_n$, зависящее от n параметров, то можно записать n первых интегралов

$$\frac{\partial S}{\partial \alpha_i} = \beta_i, \quad i = 1, \ldots, n,$$

канонической системы (7), которых, вообще говоря, достаточно чтобы получить решение системы уравнений Эйлера. Сформулируем точное утверждение.

Утверждение 6. (Теорема Якоби). Пусть $S = S(t, u, \alpha)$, $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathcal{R}_n$, — полный интеграл уравнения Гамильтона-Якоби, и

$$\det \left\| \frac{\partial^2 S}{\partial \alpha_i \partial \alpha_j} \right\| \neq 0.$$

Кроме того, пусть задан вектор $\beta = const \in \mathcal{R}_n$. Тогда функция $u = u(t, \alpha, \beta)$, определенная соотношением

$$\frac{\partial S(t, u, \alpha)}{\partial \alpha} = \beta$$

вместе с функцией

$$p = \frac{\partial}{\partial u} S(t, u(t, \alpha, \beta), \alpha)$$

образуют общее решение канонической системы уравнений (7).

Глава 9. Основы теории оптимального управления.

Основные понятия и постановка задачи. Задача поисков оптимальных режимов внешнего воздействия на динамические системы приводит к следующей математической задаче. Пусть эволюционные уравнения, описывающие изменения состояния объекта с течением времени заданы в виде дифференциальных уравнений вида

$$\frac{dx_i}{dt} = f_i(x_1, x_2, \ldots, x_n, u_1, \ldots, u_k), \quad i = 1, \ldots, n, \tag{1}$$

где $(x_1(t), x_2(t), \ldots, x_n(t)) = x(t) \in \mathcal{R}_n$ и $(u_1(t), u_2(t), \ldots, u_k(t)) = u(t) \in \mathcal{R}_k$— векторнозначные функции времени t, $t \in [t_0, t_1]$, $(f_1(x,u), \ldots, f_n(x,u)) = f(x,u) \in \mathcal{R}_n$, а $f_i(x,u)$, $i = 1, \ldots, n$ — функции, определенные и непрерывные для всех $x \in \mathcal{R}_n$ и всех $u \in \mathcal{U} \subset \mathcal{R}_k$, $\mathcal{U}$ — некоторая фиксированная область в $\mathcal{R}_k$.

Задав управляющие параметры u_i, $i = 1, \ldots, k$, как функции времени $t \in [t_1, t_2]$, получим систему уравнений

$$\frac{dx_i}{dt} = f_i(x_1, x_2, \ldots, x_n, u_1(t), \ldots, u_k(t)), \quad i = 1, \ldots, n, \tag{2}$$

которая при заданном начальном значении

$$x(t_0) = x_0 \in \mathcal{R}_n \tag{3}$$

имеет некоторое решение $x(t)$. Функции $(u_1(t), x_u(t), \ldots, u_k(t)) = u(t) \in \mathcal{R}_k$, t, $t \in [t_0, t_1]$, будем называть функциями управления; задавая их и решая задачу Коши (2)-(3), получим некоторую траекторию $x_i(t)$, $t \in [t_0, t_1]$, $i = 1, \ldots, n$.

Управлением $U = (u(\cdot), t_0, t_1, x_0)$ будем называть совокупность функций $u_i(t)$, $i = 1, \ldots, k$, заданных на отрезке $[t_0, t_1]$, и начального значения $x_0 \in \mathcal{R}_n$.

Пусть задана функция $f_0(x_1, \ldots, x_n, u_1, \ldots, u_k)$, определенная для всех $(x_1, x_2, \ldots, x_n) \in \mathcal{R}_n$ и для всех $(u_1, u_2, \ldots, u_k) = u(t) \in \mathcal{U}$, имеющая частные производные $\frac{\partial f_0}{\partial x_j}$ по переменным x_j, $j = 1, 2, \ldots, n$. Каждому управлению $U = (u(\cdot), t_0, t_1, x_0)$ поставим в соответствие число

$$J(U) = \int_{t_0}^{t_1} f_0(x, u)dt, \tag{4}$$

то есть $J(\cdot)$ — функционал, заданный на множестве управлений. Управление $U_* = (u_*(\cdot), t_0, t_1, x_0)$ назовем оптимальным, а соответствующую ему траекторию $x_*(t)$ — оптимальной траекторией, если, каково бы ни было управление $U = (u(\cdot), t_0, t_1, x_0)$, переводящее заданную точку x_0 в момент времени t_0 в точку $x_1 \in \mathcal{R}_n$ в момент времени t_1, такую, что $x_1 = x_*(t_1)$, для него выполнено неравенство

$$J(U_*) \leq J(U).$$

Будем полагать, что класс всех возможных управлений задан как класс всех кусочно непрерывных функций с разрывами первого рода, в каждый момент времени $t \in [t_0, t_1]$ принимающих значения из заданной области $\mathcal{U} \subset \mathcal{R}_k$. В частности, если

$$J(U) = \int_{t_0}^{t_1} dt = t_1 - t_0,$$

то оптимальность означает переход из точки x_0 в x_1 за минимальное время.

Задачи оптимального управления тесно связаны с задачами вариационного исчисления, рассмотренными в предыдущей главе: так как функционал (4) можно рассматривать как функционал, определенный на $(n + k)$ функциях $x_1(t), \ldots, x_n(t), u_1(t), \ldots, u_k(t)$, заданных на отрезке $[t_0, t_1]$, то есть как функционал, определенный на некотором классе кривых в $n+k+1$-мерном пространстве $\mathcal{R} = [t_0, t_1] \otimes \mathcal{R}_n \otimes \mathcal{U}$, где функции $x_1(t), \ldots, x_n(t)$ и $u_1(t), \ldots, u_k(t)$ связаны уравнениями (2), то задача оптимального управления может рассматриваться как вариационная задача на условный экстремум. Граничные условия, состоящие в том, что искомая оптимальная траектория начинается в точке x_0 и заканчивается в точке x_1, означают, что левый конец траектории принадлежит k-мерному подмножеству $\mathcal{R}$ вида $\{(t, x, u) : \ t = t_0, \ x = x_0, \ u \in \mathcal{U}\}$; а правый —$k$-мерному подмножеству $\mathcal{R}$ вида $\{(t, x, u) : \ t = t_1, \ x = x_1, \ u \in \mathcal{U}\}$.

Заметим, что простейшая вариационная задача в n-мерном пространстве $\mathcal{R}_n$, в котором подынтегральное выражение не зависит явно от времени, является частным случаем задачи об оптимальном управлении. Действительно, пусть задан функционал

$$\int_{t_0}^{t_1} f_0\left(x_1, \ldots, \mathbf{x}_n, \frac{dx_1}{dt}, \ldots, \frac{dx_n}{dt}\right) dt$$

и требуется среди кривых, проходящих через точки $x_0 \in \mathcal{R}_n$ в момент времени t_0 и $x_1 \in \mathcal{R}_n$ в момент времени t_1 найти ту, на которой этот функционал достигает минимума.

Действительно, переход к задаче оптимального управления осуществляется записью (5) в виде

$$\int_{t_0}^{t_1} f_0(x_1, \ldots, \mathbf{x}_n, u_1, \ldots, u_n) dt,$$

где функции $u_i(\cdot)$ связаны с $x_i(\cdot)$ дифференциальными уравнениями

$$\frac{dx_i}{dt} = u_i, \quad i = 1, \ldots, n.$$

Принцип максимума. Сформулируем необходимые условия минимума функционала (4), или, иначе говоря, условия, необходимые для оптимальности

управления. Для этого присоединим к системе из n уравнений (2) формально еще одно уравнение

$$\frac{dx_0}{dt} = f_0(x,u), \quad x = (x_1,\dots,x_n), \quad u = (u_1,\dots,u_k),$$

где $x_0 \in \mathcal{R}_1$ — новая формальная переменная, а $f_0(\cdot,\cdot)$ — подынтегральная функция, определяющая минимизируемый функционал (4). Одновременно начальные условия

$$x_i(t_0) = x_{i,0}, \quad i = 1,\dots,n, \tag{6}$$

дополним условием

$$x_0(t_0) = x_{0,0}. \tag{7}$$

Если U — некоторое допустимое управление, и $x = x(t)$, $t \in [t_0,t_1]$, — решение системы

$$\frac{dx_i}{dt} = f_i(x,u), \quad i = 0,1,\dots,n, \tag{8}$$

отвечающее этому управлению и начальным условиям (6)-(7), то

$$J(U) = \int_{t_0}^{t_1} f_0(x,u)dt = x_0(t_1),$$

и задача об оптимальном управлении может быть сформулирована так: найти допустимое управление U, при котором решение $x(\cdot)$ системы (8), удовлетворяющее начальным условиям (6)-(7), давало бы возможно меньшее значение $x_0(t_1)$.

Рассмотрим наряду с переменными $x_0, x_1, \dots, x_n$, новые переменные $\psi = \psi_0, \psi_1, \dots, \psi_n$, которые будем считать подчиненными уравнениям

$$\frac{d\psi_i}{dt} = -\sum_{l=0}^{n} \frac{\partial f_l}{\partial x_i}\psi_l, \quad i = 0,1,\dots,n. \tag{9}$$

Систему (9) назовем системой, сопряженной с системой (8).

Заметим, что в (9) $\frac{d\psi_0}{dt} = 0$, то есть ψ_0 — некоторая константа, не зависящая от t.

Поясним смысл введения новых переменных ψ. Определим функцию

$$\Pi(\psi,x,u) = \sum_{j=0}^{n} \psi_j f_j(x,u),$$

которую будем называть функцией Гамильтона-Понтрягина. Тогда система уравнений (9) может быть задана в виде

$$\dot{\psi}_i = -\frac{\partial \Pi(\psi,x,u)}{\partial x_i}\Big|_{\substack{u=u(t)\\ x=x(t)}}, \quad i = 0,1,\dots,n, \tag{10}$$

а систему уравнений (8) — в виде

$$\dot{x}_i = \frac{\partial \Pi(\psi, x, u)}{\partial \psi_i}\Big|_{\substack{u=u(t)\\ x=x(t)}}, \quad i = 0, 1, \ldots, n. \tag{11}$$

Эти соотношения похожи на канонические уравнения Гамильтона, рассмотренные в предыдущей главе и представляющие собой уравнения Эйлера некоторой вариационной задачи. Однако уравнения (10)-(11) не замкнуты, так как число переменных x, u, ψ превосходит число уравнений. Формулируемый ниже принцип максимума Понтрягина говорит о том, какими условиями следует дополнить уравнения (10)-(11) для получения необходимых условий оптимальности управления.

Утверждение 1. Пусть $U = (u(\cdot), t_0, t_1, x_0)$ — такое допустимое управление, что отвечающая ему интегральная кривая $x(t)$, $t \in [t_0, t_1]$, системы (8) удовлетворяет условиям

$$x_0(t_0) = 0, \quad x_i(t_0) = x_{0,i}; \quad x_i(t_1) = x_{1,i}, \quad i = 1, \ldots, n.$$

Если U — оптимально, то существуют непрерывные функции $\psi(t)$, $i = 1, \ldots, n$, $t \in [t_0, t_1]$, такие, что

1. выполнены неравенства $\psi_0 \le 0$, и $|\psi_0| + \max\limits_{i=1,\ldots,n} |\psi_i(t)| \neq 0$ для всех $t \in [t_0, t_1]$;

2. функции $\psi_i(\cdot)$, $i = 1, \ldots, n$, являются решениями сопряженной системы уравнений (9), соответствующей рассматриваемому решению $(x(\cdot), u(\cdot))$;

3. при каждом $t \in [t_0, t_1]$ функция $\Pi(\psi(t), x(t), u)$ переменной $u = (u_1, \ldots, u_k) \in \mathcal{R}_k$ достигает максимума при $u = u(t)$.

Принцип максимума может быть использован для построения оптимального управления следующим образом. Выберем для каждых фиксированных ψ и x то значение u, при котором выражение

$$\sum_{l=0}^{n} \psi_i f_l(x, u)$$

достигает максимума. Если этим условием функция u определена однозначно как функция переменных ψ и x, $u = u(\psi, x)$, то, подставив ее в (8)-(9), получим замкнутую систему $2(n+1)$ уравнений с $2(n+1)$ неизвестными, которая и определяет оптимальную траекторию.

Замечание 1. Отметим еще раз, что с помощью функции Понтрягина система уравнений (8)-(9) может быть записана в виде системы уравнений, напоминающих систему канонических уравнений Гамильтона:

$$\frac{dx_i}{dt} = \frac{\partial \Pi}{\partial \psi_i}, \quad \frac{d\psi_i}{dt} = -\frac{\partial \Pi}{\partial x_i}, \quad i = 1, \ldots, n \tag{12}$$

Однако эти уравнения имеют иной смысл: уравнения Гамильтона образуют замкнутую систему, здесь же уравнения (12) содержат, помимо $x \in \mathcal{R}_{n+1}$, $\psi \in \mathcal{R}_{n+1}$,

неизвестные величины $u \in \mathcal{R}_k$, и превращаются в замкнутую систему лишь при определенном выборе этих величин. Согласно принципу максимума, для того, чтобы в задаче об оптимальном управлении получить уравнения типа канонических, их следует записать не с помощью функции $\Pi(\psi, x, u)$, а с помощью функции

$$H(\psi, x) = \sup_u \Pi(\psi, x, u).$$

Замечание 2. Как упоминалось в предыдущей главе, динамический процесс можно рассматривать либо в терминах траекторий, которые удовлетворяют каноническим уравнениям, либо в терминах фронта волны, который удовлетворяет уравнению Гамильтона-Якоби. Подход к задачам оптимального управления, аналогичный распространению фронта волны, развивался в работах Р.Беллмана по динамическому программированию.

Рассмотрим несколько примеров.

Пример 1. Пусть требуется минимизировать функционал

$$J(u) = \int_0^T \left(u^2(t) + x^2(t)\right) dt$$

при условиях

$$\dot{x}(t) = u(t), \quad 0 \le t \le T, \qquad x(0) = x(T) = 0.$$

Очевидное решение $x(t) = 0$, $u(t) = 0$ получим, воспользовавшись принципом максимума. Функция Гамильтона-Понтрягина имеет вид

$$\Pi(\psi, x, u) = \psi_0(x^2 + u^2) + \psi u,$$

а сопряженные уравненХя — соответственно вид

$$\psi_0 = const, \quad \dot{\psi} = -\Pi_x \equiv -2\psi_0 x.$$

Если $\psi_0 = 0$, то функция $\Pi(\psi, x, u) = \psi u$ и может достигать своей точной верхней грани на всей числовой прямой только при $\psi = 0$, однако требование одновременного равенства нулю и ψ_0, и ψ противоречит условию 1 Утверждения 1. Следовательно, $\psi_0 < 0$. Положим, не ограничивая общности, $\psi_0 = -1$, тогда $\Pi(\psi, x, u) = -u^2 - x^1 - \psi u$ и достигает своей точной верхней грани по u на $\mathcal{R}_1$ при $u = \psi/2$. Тогда краевая задача принципа максимума запишется в виде

$$\dot{x} = \psi/2, \quad \dot{\psi} = 2x, \qquad 0 \le t \le T, \quad x(0) = x(T) = 0,$$

откуда однозначно определяются $x(t) = \psi(t) = 0$, $u(t) = \psi(t)/2 = 0$, $t \in [0, T]$.

Пример 2. Пусть

$$J(u) = \int_0^T \left(u^2(t) - x^2(t)\right) dt$$

при условиях

$$\dot{x}(t) = u(t), \quad 0 \le t \le T, \qquad x(0) = x(T) = 0.$$

Рассмотрим точно так же, как и в примере 1, функцию Гамильтона-Понтрягина

$$\Pi(\psi, x, u) = \psi_0(u^2 - x^2) + \psi u,$$

и сопряженные уравнения

$$\psi_0 = const, \quad \dot{\psi} = 2\psi_0 x,$$

и точно так же, как и в примере 1, убедимся, что $\psi_0 < 0$. Положим $\psi_0 = -1$ и прийдем к краевой задаче

$$\dot{x} = \psi/2, \quad \dot{\psi} = -2x, \qquad 0 \le t \le T, \quad x(0) = x(T) = 0,$$

Общее решение получившейся системы уравнений дается функциями

$$x(t) = C\sin t + D\cos t,$$

$$\psi(t) = 2C\cos t - 2D\sin t,$$

где C и D — константы. Из граничных условий при $t = 0$ найдем, что $D = 0$, а условие $x(T) = 0$ приводит к соотношению для C вида

$$C\sin T = 0.$$

Если $T \neq \pi k$, где k — положительное целое число, то $C = 0$ и оптимальное управление $u = 0$. Если же $T = \pi k$ при некотором целом положительном k, то управлений, подозрительных на оптимальные, будет бесконечно много. Это — управления вида $u = C\cos t$, и им соответствуют траектории $x(t) = C\sin t$, $t \in [0, T]$.

Данный пример интересен тем, что при $T > \pi$ существует последовательность управлений $\{u_m(t)\}$ и соответствующая ей последовательность траекторий $\{x_m(t)\}$, $t \in [0, T]$, таких, что $J(u_m) \to -\infty$ при $m \to \infty$, то есть задача оптимального управления не имеет решения, хотя краевая задача принципа максимума и разрешима. Примером таких последовательностей являются $u_m = u_m(t) = \frac{\pi m}{T}\cos\frac{\pi t}{T}$, $x_m = x_m(t) = m\sin\frac{\pi t}{T}$; $t \in [0, T]$. Действительно, тогда

$$J(u_m) = \int_0^T \left(u_m^2(t) - x_m^2(t)\right) dt = \frac{1}{2}Tm^2\left(\frac{\pi^2}{T^2} - 1\right) \to \infty \quad \text{при} \quad m \to \infty,$$

если $T > \pi$. Заметим, что последовательность $\{u_m(t)\}$ не имеет кусочно непрерывного предела.

Глава 10. Динамическое программирование.

Рассмотрим задачу оптимального управления как задачу минимизации функционала

$$J(X_0, u) = \int_{t_0}^{T} f_0(x(t), u(t), t)dt + \Phi(x(T)) \tag{1}$$

по переменным $X_0 \in \mathcal{R}_n$ и $u(\cdot)$ при дополнительных условиях

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_0 \le t \le T; \qquad x(t_0) = X_0, \tag{2}$$

где при каждом фиксированном $t \in [t_0, T]$ справедливы включения

$$x(t) \in G(t) \subset \mathcal{R}_n, \qquad u(t) \in \mathcal{U}(t) \subset \mathcal{R}_k,$$

причем $u(\cdot)$ — кусочно непрерывна, а моменты времени t_0 и T фиксированы. Задачи такого типа возникают, когда и левый, и правый концы траектории не фиксированы, и для любого $x \in \mathcal{R}_n$ значение функции $\Phi(x)$ показывает, насколько "желательно"оказаться в точке $x \in G(T) \subset \mathcal{R}_n$ в момент времени T; управление ищется из условия компромисса между величиной первого (интегрального) слагаемого в (1) и "веса"$\Phi(x)$ конечной точки траектории, $x \in G(T)$. Минимизируется функционал и выбором начальной точки траектории $X_0 \in G(t_0) \subset \mathcal{R}_n$.

Для приближенного решения задачи оптимального управления отрезок $[t_0, T]$ разбивается на N частей точками $t_0 < t_1 < \ldots < t_n = T$. Заменяя интеграл в (1) его приближенным значением по формуле прямоугольников, а дифференциальные уравнения (2) — разностными по формулам метода Эйлера, прийдем к дискретной задаче оптимального управления, в которой требуется минимизировать функцию

$$J_0(X_0, u_0, u_1, \ldots, u_{N-1}) = \sum_{i=0}^{N-1} F_i^{(0)}(x_i, u_i) + \Phi(x_N) \tag{3}$$

при дополнительных условиях

$$x_{i+1} = F_i(x_i, u_i); \quad i = 0, 1, \ldots, N-1, \quad x_0 = X_0 \in G_0 = G(t_0) \subset \mathcal{R}_n, \tag{4}$$

$$x_i \in G_i = G(t_i) \subset \mathcal{R}_n, \;\; i = 1, \ldots, N; \tag{5}$$

$$u_i \in \mathcal{U}_i = \mathcal{U}(t_i) \subset \mathcal{R}_k, \;\; i = 0, \ldots, N-1; \tag{6}$$

здесь

$$F_i^{(0)}(x, u) = f_0(x, u, t_i)(t_{i+1} - t_i) \in \mathcal{R}_1, \quad i = 0, 1, \ldots, N-1;$$
$$F_i(x, u) = x + f(x, u, t_i)(t_{i+1} - t_i) \in \mathcal{R}_n, \quad i = 0, 1, \ldots, N.$$

Заметим, что эта задача имеет и самостоятельный интерес как задача дискретного оптимального управления, в которой как координаты системы, так и сигналы управления меняются скачком в некоторые фиксированные моменты времени.

Если задать управление, то есть N векторов $u_i \in \mathcal{U}_i \subset \mathcal{R}_k$, $i = 0, 1, \ldots, N-1$, и начальное условие $x_0 = X_0 \in G_0$, то дискретная траектория $x_i \in G_i \subset R_n$, $i = 1, \ldots, N$, будет однозначно определена. Зафиксируем $x_0 = X_0$ и обозначим $\Delta_0(X_0)$ множество управлений, для которых

1. выполнены условия (6);

2. для соответствующих им траекторий выполнены соотношения (4)-(5);

множество начальных условий, для которых существует хотя бы одно управление, обозначим $\mathbf{X}_0$:

$$\mathbf{X}_0 = \{x \in G_0 : \quad \Delta_0(x) \neq \emptyset\}.$$

Тогда задача оптимального управления сводится к задаче минимизации функции многих переменных:

$$J_{0*} = \inf_{X_0 \in \mathbf{X}_0} \quad \inf_{(u_0, u_1, \ldots, u_{N-1}) \in \Delta(X_0)} J_0(X_0, u_0, u_1, \ldots, u_{N-1}).$$

Такая задача весьма громоздка, во-первых, из-за большого числа переменных — их число равно $n + Nk$, так как аргументами функции $J_0(\cdot, \cdot, \ldots, \cdot)$ являются вектор $X_0 \in \mathcal{R}_n$, и N векторов $u_i \in \mathcal{R}_k$, $i = 0, \ldots, N-1$, а во-вторых, потому, что множества $\mathbf{X}_0$ и $\Delta(X_0)$, на которых ищется минимум функционала (3), заданы неявно.

На практике для минимизации функционала (3) используется метод динамического программирования, идея которого состоит в следующем. Пусть минимум функционала (1) достигается при оптимальном управлении $u^*(t)$ на соответствующей ему оптимальной траектории $x^*(t)$, $t \in [t_0, T]$. Разобьем отрезок времени $[t_0, T]$ на два отрезка $[t_0, t_1]$ и $[t_1, T]$, и пусть известно, в чем состоит оптимальное управление на отрезке $[t_1, T]$ для любой начальной точки $X_1 \in G(t_1) \subset \mathcal{R}_n$, и каким значением функционала, определяющего качество управления, оно сопровождается. Тогда оптимальное управление на отрезке $[t_0, T]$ можно получить, рассматривая последовательно управление на отрезке $[t_0, t_1]$ и от точки X_0 в момент времени t_0 до точки X_1 в момент времени t_1, и управления на отрезке $[t_1, T]$, переводящее траекторию из X_1 в $x(T)$. Выбор оптимального управления на целом отрезке $[t_0, T]$ можно теперь осуществить, суммируя значения функционалов, определяющих качество управления от точки X_0 до X_1 и от X_1 до конечной точки траектории. Тем самым исходная задача заменяется последовательным решением двух более простых.

Чтобы осуществить эту идею на практике, рассмотрим следующие формальные понятия. Введем функцию

$$J_k(X, u_k, u_{k+1}, \ldots, u_{N-1}) = \sum_{i=k}^{N-1} F_i^{(0)}(x_i, u_i) + \Phi(x_N),$$

определенную при дополнительных условиях на x_i:

$$x_{i+1} = F_i(x_i, u_i); \quad i = k, k+1, \ldots, N-1, \quad x_k = X \in G_k \subset \mathcal{R}_n, \tag{7}$$

$$x_i \in G_i \subset \mathcal{R}_n, \;\; i = k+1, \ldots, N; \tag{8}$$

причем

$$u_i \in \mathcal{U}_i \subset \mathcal{R}_k, \;\; i = k, \ldots, N-1, \tag{9}$$

и для каждого $k = 0, 1, \ldots, N-1$ определим функцию Беллмана

$$B_k(X) = \inf_{(u_k, \ldots, u_{N-1}) \in \Delta_k(X)} J_k(X, u_k, \ldots, u_{N-1}); \tag{10}$$

здесь $\Delta_k(X)$ — множество управлений $\{u_k, u_{k+1}, \ldots, u_{N-1}\}$, для которых

1. выполнены условия (9),

2. для соответствующих им траекторий выполнены соотношения (7)-(8).

Оказывается, что функции Беллмана связаны рекуррентным соотношением

$$B_k(X) = \inf_{u \in D_k(X)} \left\{ F_k^{(0)}(X, u) + B_{k+1}(F_k(X, u)) \right\}, \quad k = 0, \ldots, N-1, \tag{11}$$

с граничным условием

$$B_N(X) = \Phi(X), \qquad X \in G_N.$$

В (11) $D_k(X)$ — множество тех $u \in \mathcal{U}_k$, для которых существует хотя бы одно допустимое управление $(u_k, \ldots, u_{N-1}) \in \Delta_k(X)$ с вектором u_k, равным u; или, иными словами, это множество значений первого вектора u_k всех допустимых управлений из $\Delta_k(X)$.

Предположим, что в результате решения рекуррентных уравнений найдены все функции Беллмана B_k, $k = 0, \ldots, N$; тогда оптимальная траектория $(x_0^*, x_1^*, \ldots, x_N^*)$ определяется следующим образом. Сначала вектор $x_0^* \in G_0$ определяется из условия

$$\inf_{x \in \mathbf{X}_0} B_0(x) = B_0(x_0^*), \tag{12}$$

а затем определяются все остальные векторы $x_i^* \in G_i$, $i = 1, \ldots, N$, последовательно полагая

$$\begin{gathered} u_0^* = u_0(x_0^*), \quad x_1^* = F_0(x_0^*, u_0^*), \\ u_1^* = u_1(x_1^*), \quad x_2^* = F_1(x_1^*, u_1^*), \ldots, \\ x_N^* = F_{N-1}(x_{N-1}^*, u_{N-1}^*). \end{gathered} \tag{13}$$

Результат, на котором основан метод динамического программирования, сформулируем в виде утверждения.

Утверждение 1. Пусть найдены функции Беллмана $B_k(\cdot)$, их области определения $\mathbf{X}_k$ и функции $u = u_k(x)$, $x \in \mathbf{X}_k$, $k = 0, \ldots, N$, на которых достигается точная нижняя грань в определении функций Беллмана (10); вектор x_0^* определен соотношением (12). Тогда оптимальное управление $(u_0^*, u_1^*, \ldots, u_{N-1}^*)$ и оптимальная траектория $(x_0^*, x_1^*, \ldots, x_N^*)$ определяются соотношениями (13).