

Державний ВНЗ «Національний гірничий університет»

М.М.Казачковський

ОПТИМІЗАЦІЯ ЕЛЕКТРОМЕХАНІЧНИХ СИСТЕМ

Частина 1

Методи оптимізації

Конспект лекцій

для студентів спеціальності 8.05070204

“Електромеханічні системи автоматизації та електропривод”

Дніпропетровськ

2014

Оптимізація електромеханічних систем. Ч. 1. Методи оптимізації: Конспект лекцій для студентів спеціальності 8.05070204 «Електромеханічні системи автоматизації та електропривод» / Уклад. М.М.Казачковський – Дніпропетровськ: Державний ВНЗ «Національний гірничий університет», 2014. – 44 с.

ЗМІСТ

1	ЗАГАЛЬНА ХАРАКТЕРИСТИКА ЗАДАЧ ОПТИМІЗАЦІЇ	4
2	МАТЕМАТИЧНІ АСПЕКТИ ОПТИМІЗАЦІЇ	9
	2.1 Властивості функції однієї змінної	9
	2.2 Властивості функції кількох змінних	10
	2.3 Обмеження	12
3	ОДНОВИМІРНІ МЕТОДИ ПОШУКУ ЕКСТРЕМУМУ	15
	3.1 Метод половинного ділення (дихотомії)	16
	3.2 Метод золотого перерізу	17
	3.3 Метод квадратичної апроксимації	18
	3.4 Метод Ньютона	19
4	БАГАТОВИМІРНІ МЕТОДИ ПОШУКУ	20
	4.1 Метод покоординатного пошуку Гауса-Зайделя	20
	4.2 Метод спряжених напрямів Пауела	20
	4.3 Метод Хука-Дживса (пошук за зразком)	21
	4.4 Симплексний метод	23
	4.5 Метод деформовуваних багатогранників Нелдера-Міда	25
	4.6 Метод комплексів Бокса	26
	4.7 Випадковий пошук	27
	4.8 Метод найшвидшого спуску	28
	4.9 Метод спряжених градієнтів Флетчера-Рівса	31
	4.10 Особливості оптимізації з обмеженнями	31
	4.11 Метод штрафних функцій	34
	4.12 Лінійне програмування	36
5	ЕКСТРЕМАЛЬНІ СИСТЕМИ РЕГУЛЮВАННЯ	40
	5.1 Загальна характеристика екстремальних систем регулювання	40
	5.2 Структури екстремальних систем керування	41
	5.3 Особливості багатовимірних систем екстремального керування	43
	Література	44

1 ЗАГАЛЬНА ХАРАКТЕРИСТИКА ЗАДАЧ ОПТИМІЗАЦІЇ

Пошук оптимального рішення – одна з найпоширеніших науково-технічних задач. Подібні задачі виникають під час проектування, керування динамічними системами та технологічними процесами, планування та організації виробництва, в наукових дослідженнях.

Метою пошуку є вибір таких параметрів об'єкту (системи, процесу), котрі забезпечують найкраще (максимальне або мінімальне) значення певної найбільш важливої його характеристики (властивості). У літературі цю характеристику (надалі позначатимемо її через Y) називають цільовою функцією (ЦФ), функцією цілі, функцією відгуку, оптимізовуваною функцією, залежною (керованою) змінною. Параметри, зміна яких забезпечує оптимізацію ЦФ, називають незалежними (керуючими) змінними (факторами). Будемо позначати їх як X_i ($i=1, 2, \dots, n$).

Залежно від типу об'єкту, що оптимізується, виконуваних ним функцій, а також ступеню деталізації оптимізаційної задачі в якості ЦФ можуть бути вибрані (за категоріями):

- економічні: вартість, експлуатаційні (або зведені) витрати, прибуток (збитки), кількість зайнятих працівників тощо;
- загально-технічні: надійність, витрати енергії (палива, матеріалів), вага, габарити, коефіцієнт корисної дії, продуктивність, кількість відходів (шкідливих викидів);
- спеціальні: точність, швидкодія, пройдений шлях, тривалість пуску, температура нагрівання обмоток двигуна, розмах коливань, перевантажувальна здатність, середньоквадратична похибка регулювання, коефіцієнт потужності, втрати електроенергії, споживана потужність (струм).

Цільова функція має відповідати низці вимог:

- кількісний характер;
- однозначність (заданому набору факторів відповідає з точністю до похибки експерименту або обчислень одне значення ЦФ);
- змістовність (ЦФ повинна оцінювати ефективність функціонування об'єкту у наперед заданому сенсі);
- фізичний зміст (можливість змістовної інтерпретації значення ЦФ);
- обчислюваність.

Основні вимоги до факторів:

- керованість (можливість цілеспрямованого впливу на них, некеровані фактори звичайно відносять до збурень);
- незалежність (можливість вибору рівня одного з факторів незалежно від рівня інших);
- значимість (наявність впливу з боку фактору на ЦФ)

Нерідко ефективність роботи системи оцінюють за допомогою кількох критеріїв, які часто суперечать один одному. Так, зменшення ваги механізму звичайно супроводжується зниженням його надійності, підвищення швидкодії системи керування – збільшенням коливальності тощо. Оскільки одночасно оптимізувати кілька ЦФ, що суперечать одна одній, неможливо, одну з них можна обрати як

основну, а інші виступатимуть як обмеження, що задані у вигляді рівностей або нерівностей, наприклад:

- мінімізувати масу за умови, що міцність конструкції буде не менша допустимої;
- мінімізувати втрати в асинхронному двигуні за умови, що критичний момент дорівнює заданому, а початковий пусковий не менший від допустимого;
- максимізувати продуктивність технологічного процесу за умови, що шкідливі викиди не перевищують припустимого рівня.

Подібні обмеження, котрі, як і ЦФ, є функціями незалежних змінних, називають функціональними. Причиною появи функціональних обмежень найчастіше є обмеженість ресурсів (матеріальних, фінансових, енергетичних, інформаційних тощо).

Обмеженими також можуть бути значення факторів. Причини цього найчастіше фізичні (опори, індуктивності, коефіцієнт потужності не можуть бути від'ємними, кількість працівників від'ємним або нецілим, відносний вміст компонентів у сплаві лежить у межах 0...100% тощо). Функціональні обмеження також обмежують припустимі значення факторів, проте неявним чином.

Одні й ті ж самі змінні залежно від формулювання оптимізаційної задачі можуть виступати або як цільова функція, або функціональне обмеження, або навіть як фактори. Так, в одній задачі мінімізується чисельність працівників (це – ЦФ) шляхом раціональної організації їхньої праці, в іншій – мінімізуються зведені витрати шляхом вибору оптимальної кількості зайнятих (це – фактор). Змінюючи активний опір ротора асинхронного двигуна, що працює у повторно-короткочасному режимі з частими пусками, можна мінімізувати середні втрати у двигуні (це – ЦФ) за пускового моменту, не меншого за допустимого (це – функціональне обмеження). В іншій ситуації треба максимізувати пусковий момент (ЦФ) за умови, що середні втрати не перевищують номінальних (функціональне обмеження).

Для обґрунтованого вибору ЦФ необхідне знання технологічних особливостей оптимізованого об'єкту, вимог до нього. Умови, в яких експлуатується об'єкт, нерідко диктують суттєво різні критерії оптимізації. Так, оптимізуючи електрообладнання, що живиться від автономних (тобто відносно малопотужних) джерел (самохідні машини, літальні апарати, судна, тепловози, годинники, мобільні телефони), нерідко обирають споживану потужність як основну ЦФ. Основні ж вимоги до електроприводів верстатів з числовим програмним керуванням – точність та швидкодія. Помилки у виборі ЦФ неминуче призведуть до невірної результату. Так, якщо під час вибору місця розташування підстанції керуватися критерієм мінімальної сумарної довжини відхідних ліній, шуканою точкою стане геометричний центр споживачів, а обравши критерієм мінімізацію втрат в лініях, отримаємо т.зв. центр навантажень, зміщений у бік більш потужних споживачів.

На вибір ЦФ та факторів впливає також ступінь деталізації задачі. Під час розробки складних систем (технологічний процес, загальнопромислова серія електричних двигунів, агрегат високої вартості) часто використовують економічні та

техніко-економічні критерії, На стадії деталізації (розробка окремих вузлів агрегату, спеціального двигуна для конкретного механізму) на перший план виступають технічні критерії (точність, вага, температура нагріву обмоток двигуна тощо).

Класифікація задач оптимізації.

1. За фізичною сутністю оптимізованого об'єкту:
 - натурний (діючий) об'єкт;
 - модель об'єкту (фізична або математична).
2. За способом розв'язання задачі:
 - аналітичні (на основі алгебраїчних співвідношень або з використанням диференціального або варіаційного числення);
 - графічні (звичайно за кількості факторів, не більших двох);
 - чисельні (завжди ітераційні).
3. За характером задачі та отриманого результату:
 - пошук оптимальних параметрів (результат – числа);
 - пошук оптимальних функцій (найчастіше результат – функція часу).

Хоча під час експериментування на діючому об'єкті часто може бути отримана найбільш достовірна інформація, можуть виникнути певні проблеми:

- вимірювання незалежних та залежних змінних з похибками внаслідок впливу випадкових збурень;
- зміщення положення оптимуму у часі внаслідок впливу неврахованих або неконтрольованих систематичних збурень;
- дорожнеча або небезпечність експериментів;
- недоступність і навіть відсутність натурального об'єкту.

Іноді, коли отримання математичного опису об'єкту з певних причин неможливе, натурний експеримент є єдиним вибором. Проте у більшості випадків дослідження здійснюють за допомогою більш-менш достовірного математичного опису (моделі). Особливою формою оптимізації натурального об'єкту можна вважати процес самонастроювання цифрових регуляторів сучасних електроприводів з цифровим керуванням, а також процес пошуку точки максимуму вихідної потужності сонячної батареї.

Застосування аналітичних методів передбачає можливість перетворення математичної моделі до такого вигляду, що дозволяє використання стандартних процедур математичного аналізу. На жаль, у більшості технічних задач це неможливо, оскільки модель нерідко буває заданою у вигляді системи алгебраїчних або диференціальних рівнянь (часто нелінійних), які не розв'язуються у загальному вигляді. Тому неминуче використання чисельних методів оптимізації, вивчення яких і є основною задачею даного курсу. Слід зазначити, що ці методи можуть успішно використовуватися і для оптимізації натурних об'єктів та математичних моделей.

Чисельні методи пошуку оптимальних параметрів є предметом вивчення такої галузі знань, як математичне програмування (МП). У ході розвитку МП виділилися наступні його основні розділи.

- Лінійне програмування (ЛП). Використовується для задач, у яких ЦФ та обмеження є лінійними функціями факторів:

$$Y = \sum_{i=1}^n c_i X_i \rightarrow \min;$$

$$G = \sum_{j=1}^m a_j X_j \leq b.$$

Найбільшого застосування ЛП знаходить у задачах з економіки, фінансів, організації виробництва тощо.

- Квадратичне програмування. Розроблено для задач, у яких ЦФ та обмеження, окрім лінійних членів $c_i X_i$, залежать від добутків факторів:

$$Y = \sum_{i=1}^n c_i X_i + \sum_{\substack{i=1; \\ j=1}}^n d_{ij} X_i X_j \rightarrow \min .$$

- Геометричне програмування. Передбачає, що ЦФ та обмеження задані як суми добутків факторів, піднесених до певного ступеня:

$$Y = \sum_{j=1}^m C_j \prod_{i=1}^n X_i^{a_{ji}} .$$

- Нелінійне програмування (НЛП), яке призначене для задач, у яких ЦФ та (або) хоча б одне з обмежень пов'язані з факторами нелінійними залежностями загального вигляду. Найчастіше такі задачі трапляються в техніці (особливо, коли оптимізовуваний об'єкт описується системою нелінійних диференціальних рівнянь). Саме методам НЛП буде присвячена більша частина даного курсу.

Задачі пошуку оптимальних функцій (найчастіше керуючих впливів оптимальної форми) характерні для теорії оптимального керування.

Розв'язання оптимізаційної задачі можна поділити на такі етапи.

1. Постановка задачі оптимізації. Задача формулюється звичайно словесно, у змістовних термінах відповідної галузі знань. На цьому етапі зазвичай обирають цільову функцію, незалежні змінні (принаймні основні), обмеження. Слід зазначити, що в технічних задачах обмеження звичайно присутні. Їхня відсутність найчастіше свідчить про те, що задачу погано сформульовано.
2. Формалізація задачі (переклад словесного формулювання на мову математики). На даному етапі записують рівняння, які пов'язують ЦФ та обмеження з незалежними змінними, визначають діапазон зміни останніх. Можливе також певне коригування переліку факторів та обмежень.
3. Вибір методу розв'язання задачі. Залежно від характеру рівнянь обирають поміж аналітичним та чисельним способом, а в останньому випадку обирають також конкретний чисельний метод пошуку.
4. Суто рішення.
5. Аналіз отриманих результатів.

До найбільш важливих слід віднести перші два етапи. Правильна постановка задачі – уже половина успіху. Навчити цього нелегко, проте навчитися можна. Нерідко одну й ту ж саму задачу може бути формалізовано різним чином. Від то-

го, наскільки вдало проведено формалізацію, може залежати не тільки швидкість пошуку, а й сама можливість досягнення екстремуму.

Важливу роль на перших етапах оптимізації відіграє апріорна інформація про оптимізовуваний об'єкт. Її можна дістати з теорії, яка описує функціонування об'єкту, результатів попередніх експериментів, моделювання, нарешті особистого досвіду. Завдяки цій інформації у певних випадках можна передбачити характер ЦФ (наявність глобального або локальних екстремумів та їхнє приблизне розташування), порівняльну значимість окремих факторів, перелік та характер обмежень тощо. Апріорна інформація може допомогти сформулювати та формалізувати задачу, обрати початкове наближення, метод розв'язання задачі.

У практичних випадках можуть стояти завдання пошуку як мінімуму, так максимуму. Проте у літературі, присвяченій оптимізації, та у відповідних комп'ютерних математичних пакетах найчастіше йдеться про пошук мінімуму, оскільки формально задача максимізації може бути елементарно зведена до задачі мінімізації шляхом зміни знаку у правій частині цільової функції.

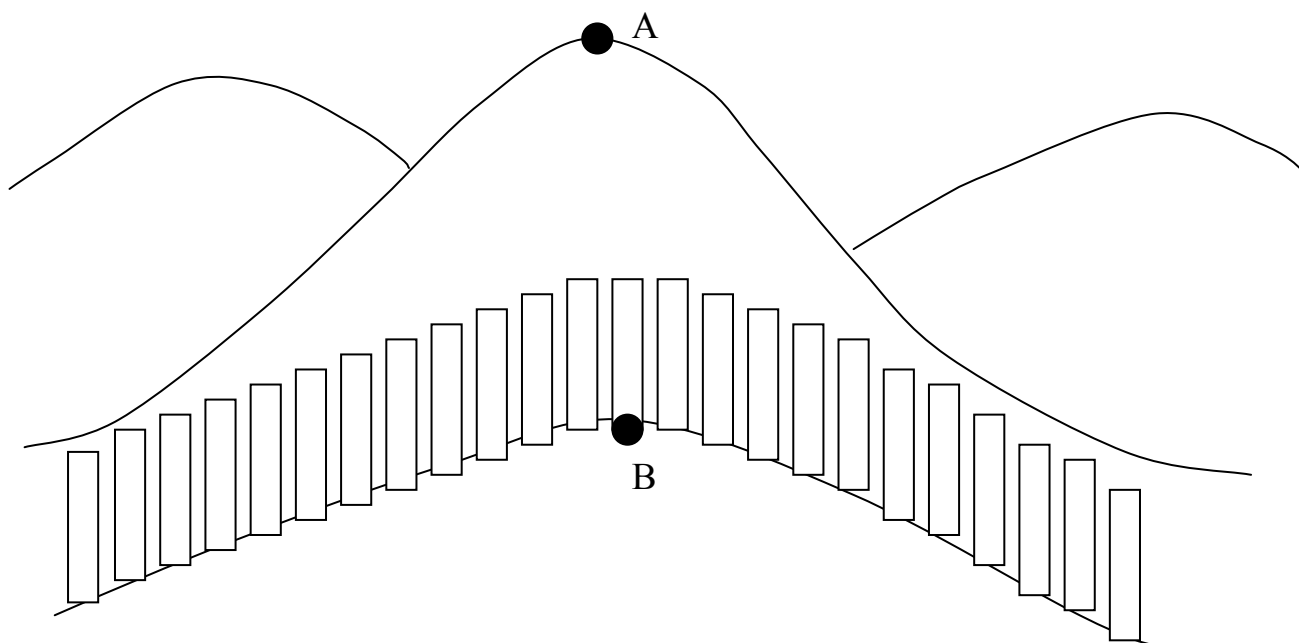


Рис. 1.1

Наочно типову оптимізаційну задачу можна проілюструвати за допомогою рис. 1.1. Найвищою точкою гори є, звичайно, т. А. Проте внаслідок обмеження (паркану) вона є недосяжною. Тому найвищою точкою з урахуванням обмеження є точка В. Задачею оптимізації і є пошук цієї точки (характер обмеження заданий).

2 МАТЕМАТИЧНІ АСПЕКТИ ОПТИМІЗАЦІЇ

У даному розділі мова йтиме про основні поняття теорії оптимізації, необхідні для розуміння методів пошуку екстремуму, які розглядатимуться у наступних розділах.

2.1 Властивості функції однієї змінної

Монотонною функцією називають таку, яка за монотонної зміни (зростання або спадання) фактору на інтервалі $X_1...X_2$ не змінює знаку своєї похідної. Це означає що на цьому інтервалі вона не має екстремуму (як усі три функції на рис. 2.1).

Немонотонною унімодальною функцією є така, що за умови монотонної зміни фактору одноразово змінює знак своєї похідної (тобто має один екстремум). Функції, графіки яких зображено на рис. 2.2, є унімодальними. Унімодальні функції є монотонними по обидва боки від свого єдиного екстремуму.

Якщо функція має кілька екстремумів, її називають мультимодальною (полімодальною). Вона має один глобальний екстремум (найбільший з максимумів або найменший із мінімумів) та хоча б один локальний. На рис. 2.3 X_{03} – глобальний мінімум, X_{04} – глобальний максимум, X_{01} та X_{05} – локальні мінімуми та X_{04} – локальний максимум.

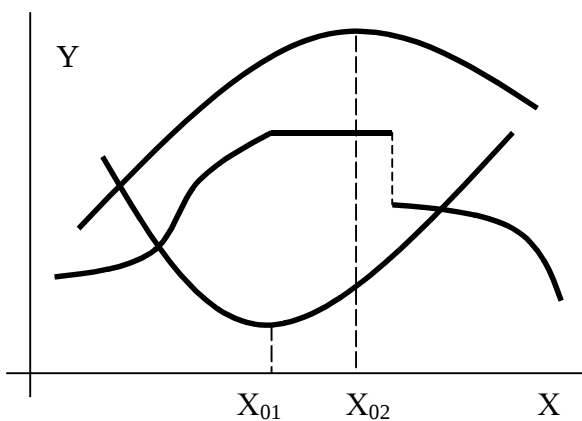


Рис. 2.2 Унімодальні функції

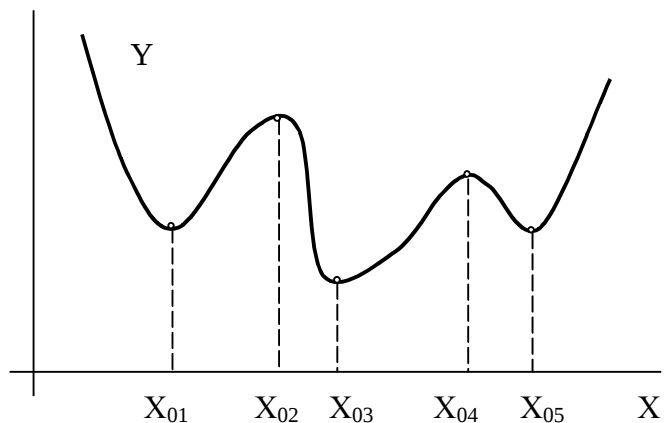


Рис. 2.3 Мультимодальна функція

Відомо, що у точці екстремум перша похідна функції змінює знак. Тому необхідною умовою наявності екстремуму є нульове значення першої похідної функції у певній точці $X=X_0$:

$$\left. \frac{dY}{dX} \right|_{X=X_0} = 0.$$

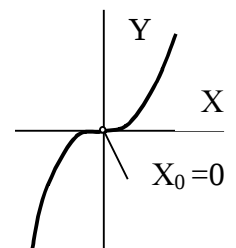


Рис. 2.4

Проте ця умова не є достатньою, оскільки вона виконується і для точки перегину (рис. 2.4). Якщо перші $n-1$ вищі похідні функції в точці $X=X_0$ дорівнюють нулю, а похідна порядку n – ненульова, достатньою умовою буде па-

рний номер ненульової похідної (якщо її номер непарний, у точці $X=X_0$ наявний перегин). Проілюструємо це на прикладі двох функцій:

$$y_1 = x^3;$$

$$\frac{dy_1}{dx} = 3x^2 = 0 \text{ у точці } x = 0.$$

У тій самій точці

$$\frac{d_2 y_1}{dx^2} = 6x = 0;$$

$$\frac{d_3 y_1}{dx^3} = 6 \neq 0.$$

Оскільки номер похідної, відмінної від нуля є непарним, у точці $x=0$ функція $y_1 = x^3$ має перегин.

$$y_2 = x^2;$$

$$\frac{dy_2}{dx} = 2x = 0 \text{ у точці } x = 0.$$

У тій самій точці

$$\frac{d_2 y_2}{dx^2} = 2 \neq 0.$$

Оскільки номер похідної, відмінної від нуля є парним, у точці $x=0$ наявний екстремум.

Неабиякого впливу на хід пошуку екстремуму мають розриви функції та їхніх похідних. На рис. 2.5 розрив має функція Y_1 , а функції Y_2 та Y_3 – розриви похідних.

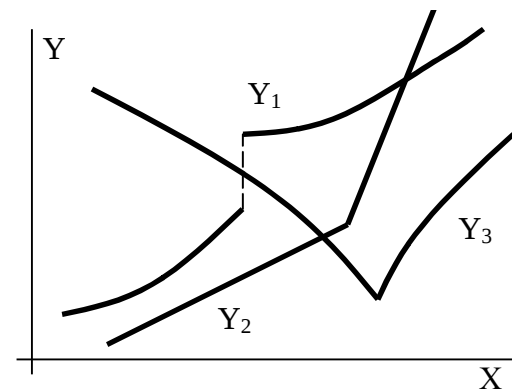


Рис. 2.5 Розриви функції та її похідних

2.2 Властивості функції кількох змінних

Функції кількох змінних (або багатовимірні функції) залежать від кількох незалежних змінних (факторів). Факторним простором називають умовний простір, координатами якого є фактори. Розмірність факторного простору дорівнює кількості факторів. Так, для функції однієї змінної факторним простором є лінія (вісь абсцис, її розмірність дорівнює одиниці), факторним простором функції двох змінних є координатна площина (розмірність – два), функція трьох змінних має факторний простір у вигляді тривимірного простору. Для задач з більшою кількістю факторів говорять про факторний гіперпростір. Надалі ми заради наочності ілюструватимемо властивості функцій кількох змінних та багатовимірних методів пошуку на прикладі функцій двох змінних.

Топографію функцій двох змінних на факторній площині зручно відображати за допомогою ліній рівня. Подібний метод широко використовують на географічних картах для відображення рельєфу місцевості. На рис. 2.6 поверхня $Y=f(X_1, X_2)$ відображає залежність функції Y від двох факторів X_1, X_2 (цю поверхню іноді називають поверхнею відгуку). Січна площина $Y=Y_1$ є паралельною факторній площині X_1OX_2 та має від неї відстань Y_1 . Лінія рівня – це проекція на факторну площину X_1OX_2 лінії, яка є перерізом поверхні відгуку та січної площини.

Інакше кажучи, лінія рівня – це геометричне місце точок на факторній площині, яким відповідає однакове значення функції. Тривимірним аналогом ліній рівня є поверхні рівня.

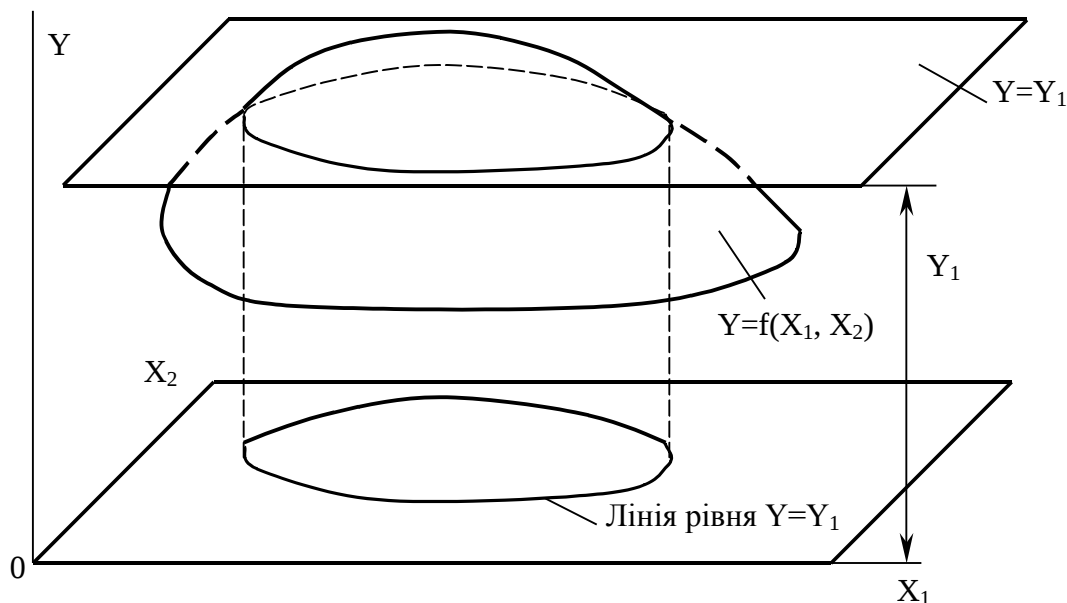


Рис. 2.6 До поняття лінії рівня

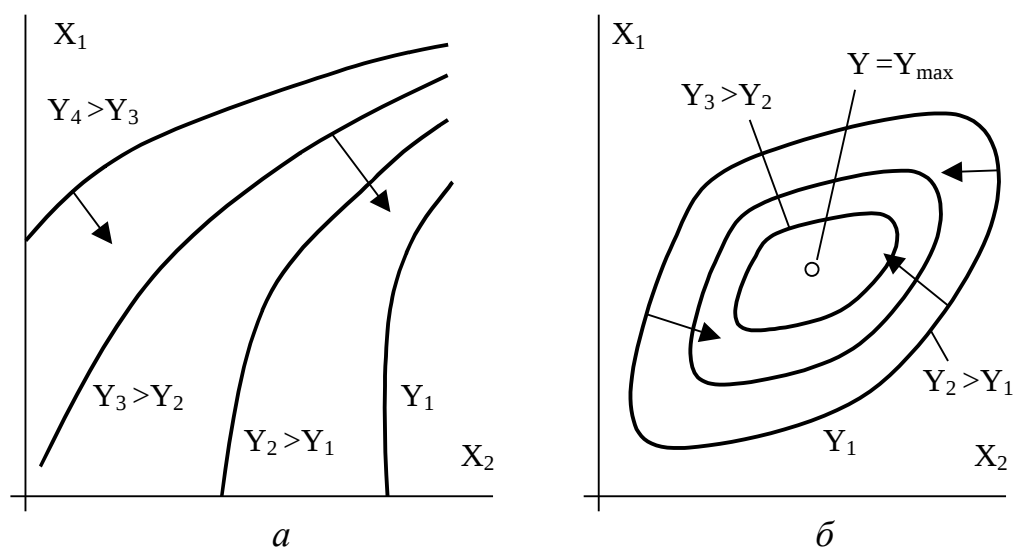


Рис. 2.7 Лінії рівня монотонної (а) та немонотонної (б) функцій двох змінних

Маючи сукупність ліній рівня, можна судити про основні властивості двовимірної функції (монотонність, опуклість, наявність екстремумів тощо). Якщо лінії рівня незамкнені (рис. 2.7, а), функція є монотонною (принаймні у локальній області, у якій лінії рівня будувалися). У разі замкнених ліній рівня (рис. 2.7, б)десь усередині цих ліній обов'язково є принаймні один екстремум (точка $Y=Y_{\max}$).

Важливим в теорії оптимізації є поняття градієнту. Градієнтом називають вектор, який показує напрям найшвидшого зменшення функції у заданій точці (вода тече за напрямом градієнту). У просторі двох факторів градієнт завжди ортогональний лінії рівня, яка проходить цією точкою. Для його визначення необхідно знати частинні похідні ЦФ за усіма факторами. На рис. 2.7 градієнти показані стрілками.

Немонотонні (екстремальні) функції можуть бути опуклими та неопуклими. Відрізнити немонотонні опуклі функції можна за допомогою наступного правила: усі їхні замкнені лінії рівня мають таку форму, що будь-які дві точки (A, B) усередині лінії рівня поєднуються відрізком прямої, яка не перетинає лінії рівня (рис. 2.8,а). Якщо ж усередині замкненої лінії рівня є такі дві точки, що відрізок, який їх з'єднує, перетинає цю лінію рівня, функція не є опуклою (рис. 2.8,б). Неопуклість функції може бути (хоча й не завжди) ознакою її мультимодальності. Так, на рис. 2.8,в зображено лінії рівня полі модальної функції, яка має два максимуми (точки C та D). Між максимумами є особлива точка H , у якій одна з ліній рівня перетинає сама себе. Якщо рухатися через неї траєкторією CHD , у цій точці спостерігатиметься мінімум, якщо траєкторією EHF – максимум. Тому подібні точки називають точкою мінімаксу або сідловиною. Географічним аналогом такої точки є гірський перевал між двома горами.

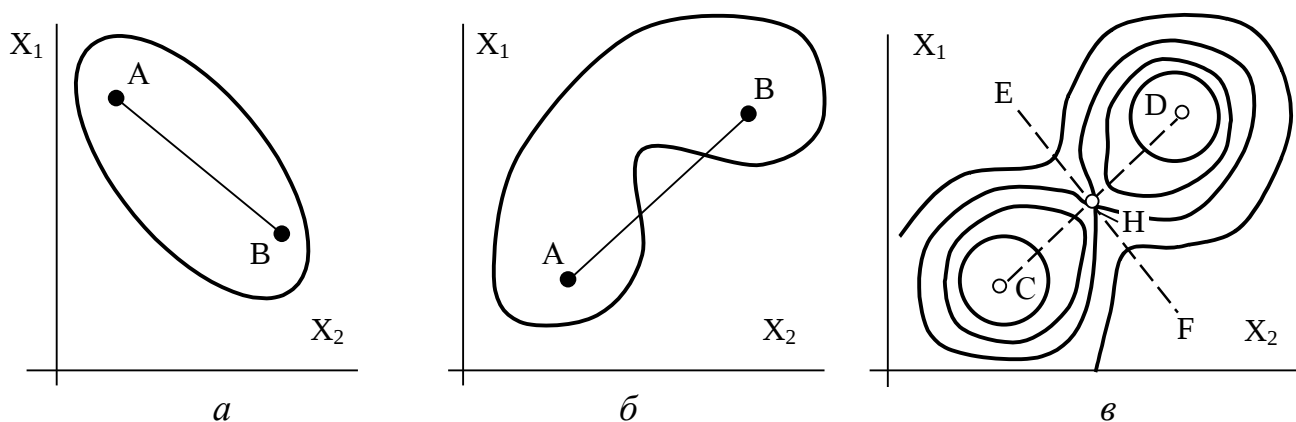


Рис. 2.8 Лінії рівня опуклої (а) та неопуклої (б, в) функції

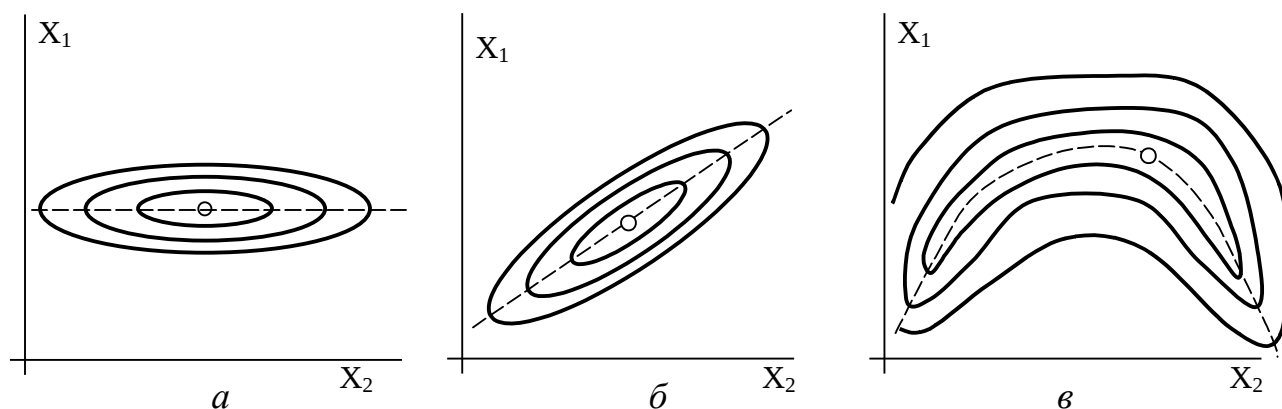


Рис. 2.9 Поверхні відгуку, які мають яри

Деякі багатовимірні функції (рис. 2.9) мають поверхню відгуку з т.зв. ярами (вузькими западинами). Чим вужчий яр та чим більше його поздовжня вісь відрізняється від прямої лінії, тим повільнішим буде процес пошуку екстремуму. З показаних на рис. 2.9 найгіршим буде варіант рис. 2.9,в.

2.3 Обмеження

Якщо на значення екстремальної функції або факторів не накладається жодних умов (обмежень), її екстремум є безумовним (як на рис. 2.2, 2.3). Обмеження поділяє область визначення функції на дозовану та заборонену області. Якщо

обмеження, яке накладається на фактори, робить безумовний екстремум недосяжним, маємо умовний екстремум на межі допустимої області (безумовний екстремум потрапив у заборонену область). Функція, зображена на рис. 2.10,а у точці $X=X_0$ має безумовний екстремум, а у точці $X=X_G$ – умовний за умови, що $X \geq X_G$. Тут і надалі штрихуванням з одного боку обмеження позначається заборонена область. Слід зазначити, що за наявності обмежень умовний екстремум може мати навіть монотонна функція, яка не має безумовного екстремуму.

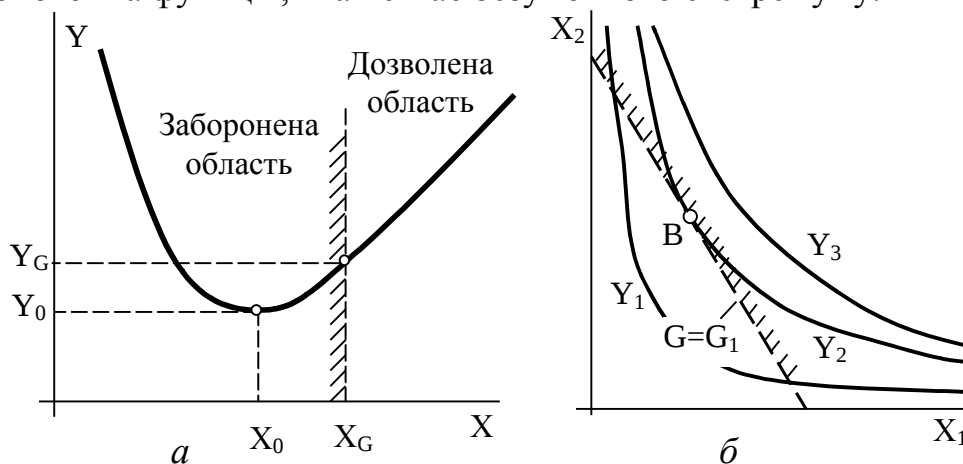


Рис. 2.10 Безумовний та умовний екстремуми

У багатofакторних задачах також трапляються обмеження на рівень факторів (на рис. 2.11,а умовним екстремумом є точка В). Окрім того, розповсюджені (особливо в технічних задачах) також функціональні обмеження, які обмежують область визначення цільової функції неявним чином. Так, на рис. 2.11,б окрім ліній рівня ЦФ Y показані також лінії рівня параметру $G=f(X_1, X_2)$, який так само є функцією тих самих факторів. Якщо задачу сформульовано таким чином, щоб знайти екстремум функції Y за умови, що функція $G \leq G_2$, то область вище та правіше лінії $G = G_2$ буде забороненою, а умовним екстремумом буде точка В.

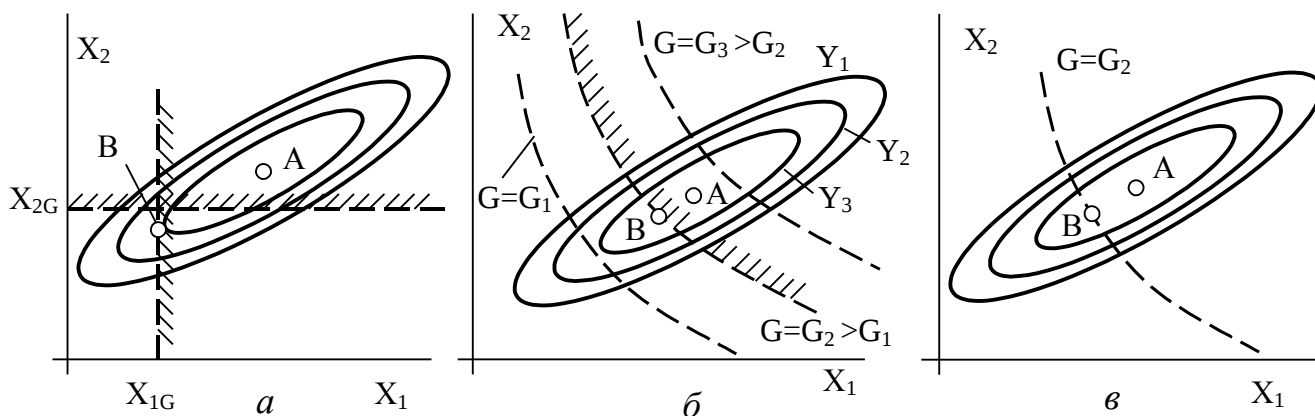


Рис. 2.11 Типи обмежень у двофакторних задачах

Розглянуті обмеження були обмеженнями типу нерівності. Якщо задачу, зображену на рис. 2.11,б, переформулювати таким чином, щоб екстремум ЦФ шукався за умови, що $G = G_2$, то забороненою областю буде вся координатна площина $X_1 O X_2$, окрім лінії $G = G_2$ (рис. 2.11,в), а умовним екстремумом – точка В. Подібні обмеження називають обмеженнями типу рівності. Оскільки обмеження

типу рівності пов'язує два фактори один з одним через рівняння обмеження $G=f(X_1, X_2)=G_2$, задача з двофакторної перетворюється на однофакторну (пошук екстремум не на площині, а на лінії $G=G_2$). Уведення кожного подібного обмеження зменшує розмірність задачі на одиницю.

У багатofакторних задачах накладення обмеження навіть на монотонну функцію може призвести до появи екстремуму (т. *B* на рис. 2.10,б).

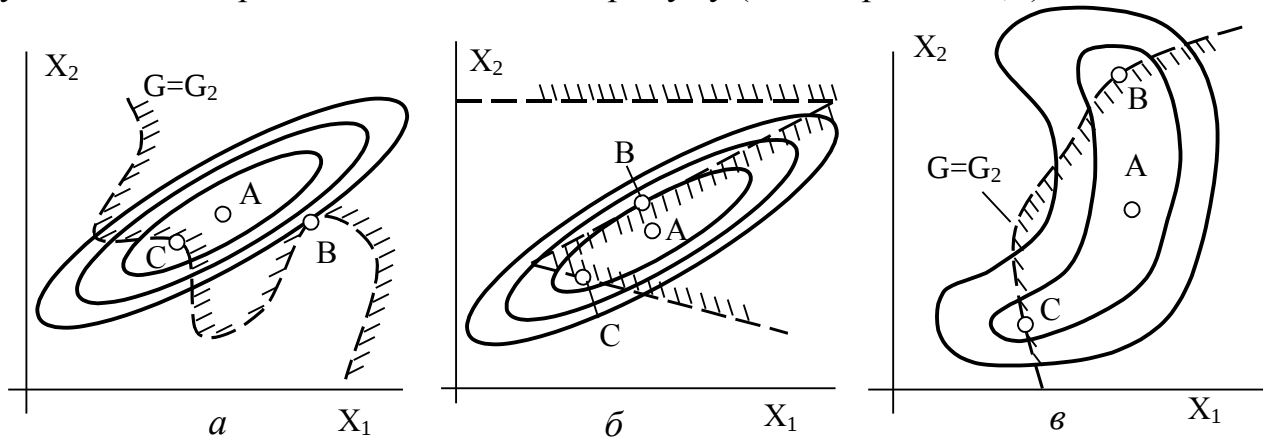


Рис. 2.12 Неопуклість як причина мультимодальності

Якщо ЦФ або обмеження є неопуклими, накладення обмеження може призвести до виникнення кількох екстремумів. Накладення неопуклих обмежень на опуклу цільову функцію (рис. 2.12,а,б) або опуклого обмеження на неопуклу (хай і унімодальну) цільову функцію (рис. 2.12,в) призвело до появи замість одного безумовного (т. *A*) двох умовних екстремумів (тт. *B* та *C*).

3 ОДНОВИМІРНІ МЕТОДИ ПОШУКУ ЕКСТРЕМУМУ

Хоча більшість практичних задач оптимізації є багатфакторними, це не зменшує актуальності одновимірних методів пошуку. Як буде показано у розділі 4, процедура багатьох багатовимірних методів зводиться до послідовності одновимірних пошуків уздовж певних прямих у факторному просторі великої розмірності.

Методи пошуку (одно- та багатовимірні) поділяють на два великих класи:

- прямі (не використовують похідну ЦФ у процесі пошуку, а порівнюють лише сусідні значення самої ЦФ);
- непрямі (для прискорення пошуку застосовують похідні ЦФ).

Прямі методи вирізняються надійністю в ситуаціях, коли ЦФ або обмеження мають розриви похідних, проте більш повільні порівняно з непрямыми.

Окрім того, методи пошуку поділяють на безумовні (не здатні врахувати наявність обмежень) та умовні (такі, що враховують обмеження).

Якість методів пошуку оцінюють звичайно за такими параметрами:

- швидкість збіжності (кількість ітерацій, потрібних для досягнення мінімуму);
- тривалість виконання однієї ітерації;
- об'єм пам'яті ЕОМ, необхідний для реалізації методу;
- клас вирішуваних задач (розмірність, лінійність або нелінійність, наявність обмежень тощо).

Одновимірний пошук складається з наступних етапів.

- Визначення напрямку руху до екстремуму. Рішення приймається на основі апріорної інформації або після серії спеціальних експериментів.
- Встановлення меж (локалізація екстремуму). Результатом цього етапу є інформація про відрізок на осі абсцис (інтервал локалізації), усередині якого достовірно існує мінімум ЦФ.
- Уточнення положення мінімуму (звуження інтервалу локалізації до наперед заданих розмірів).

Деякі одновимірні методи передбачають, що мінімум уже локалізований. У цьому випадку для локалізації можна використати евристичний метод, згідно з яким наступний крок неупинно збільшується, доки ЦФ не почне зростати (екстремум узятий у “вилку”). Величина k -го кроку

$$\Delta_k = X_k - X_{k-1} = 2^{k-1} \Delta,$$

де X_{k-1} та X_k – значення фактору до та після k -го кроку, Δ – початковий крок.

Якщо виконується умова:

$$Y(X_{k-1}) \geq Y(X_k) \leq Y(X_{k+1}),$$

екстремум локалізовано в інтервалі $[X_{k-1}; X_{k+1}]$. На рис. 3.1 показано процес локалізації починаючи з точки з абсцисою X_0 . Після третього кроку виявилось, що екстремум розташований в інтервалі $[X_1; X_3]$.

Слід зазначити, що усі методи оптимізації, розглянуті у цій та наступних розділах, є локальними (тобто здатні протягом одного пошуку знайти лише один екстремум).

3.1 Метод половинного ділення (дихотомії)

Передумовою методу є знання початкового інтервалу локалізації. Процедура пошуку складається з наступних етапів.

1. Ділення інтервалу локалізації $[a; b]$ шириною L на чотири однакових підінтервали (рис. 3.2).
2. Розрахунок значень ЦФ у трьох точках (X_1, X_0, X_m) усередині інтервалу локалізації (значення ЦФ на краях інтервалу відомі з попереднього етапу).
3. Порівняння значень отриманих ЦФ та відкидання підінтервалів, які не містять мінімуму. Ширина нового інтервалу локалізації після цього зменшується удвічі. Після виконання п.2 можливі три ситуації залежно від співвідношення отриманих значень ЦФ (рис. 3.2).
Очевидно, що результатами порівняння будуть:

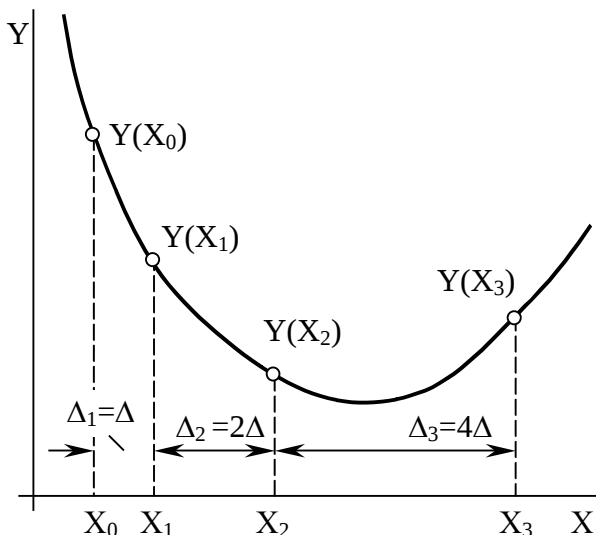


Рис. 3.1 Локалізація екстремуму евристичним методом

- а) мінімум належить діапазону $[a; X_0]$, два праві підінтервали відкидаються;
 - б) мінімум – у діапазоні $[X_1; X_2]$, відкидаються обидва крайні підінтервали;
 - в) новим інтервалом локалізації є $[X_0; b]$, обидва ліві підінтервали є зайвими.
4. Після звуження інтервалу локалізації відбувається переприсвоєння імен границь (крайнім точкам надаються імена a та b , середній – X_0). Після цього всередині отриманого інтервалу локалізації реалізують ще два експерименти, абсциси яких зсунуті усередину від країв на чверть довжини нового інтервалу локалізації ($X_1 = a + L/4$ та $X_2 = b - L/4$).
 5. Повернення до п.3.

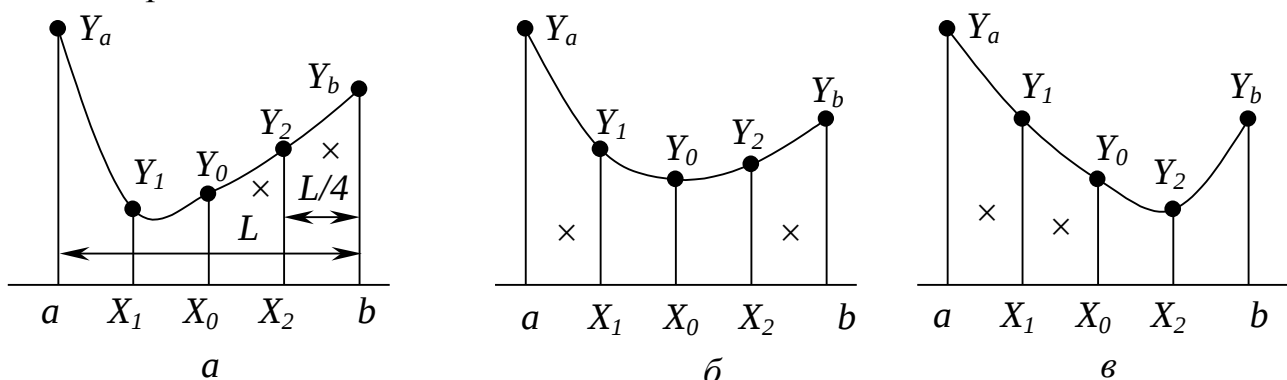


Рис. 3.2 Ситуації перед відкиданням підінтервалів

Критеріями закінчення пошуку можуть бути досягнення достатньо малої зміни ЦФ

$$\left| \frac{Y_{\max} - Y_{\min}}{Y_{\min}} \right| \leq \delta_Y \quad (3.1)$$

або достатньо вузького інтервалу локалізації

$$\left| \frac{b-a}{X_0} \right| = 2 \left| \frac{b-a}{b+a} \right| \leq \delta_x. \quad (3.2)$$

Тут максимальне $Y_{\max}$ та мінімальне $Y_{\min}$ значення ЦФ обираються з трьох останніх найкращих результатів.

У кожному циклі пошуку за методом дихотомії реалізується два додаткових експерименти.

3.2 Метод золотого перерізу

Як і для попереднього методу, початковий інтервал локалізації вважається відомим. Особливістю методу є ділення інтервалу локалізації на три нерівних частини, причому співвідношення ширин сусідніх частин, а також ширин інтервалів локалізації на суміжних циклах є незмінним. На рис. 3.3 через $L_{k-1} = b - a$ позначено довжину інтервалу локалізації на попередньому етапі, точки X_1 та X_2 зсунуті від дальніх країв цього інтервалу на однакові відстані $\tau \cdot L_{k-1}$. Після звуження інтервалу локалізації шляхом відкидання підінтервалу $[a; X_1]$ ширина нового інтервалу локалізації буде $L_k = \tau \cdot L_{k-1}$. Тут $\tau = \frac{L_k}{L_{k-1}} = \frac{L_{k+1}}{L_k} < 1$ – кратність зменшення

ширини інтервалу локалізації. Її значення можна отримати як один із коренів рівняння

$$\tau \cdot L_{k-1} + \tau^2 \cdot L_{k-1} = L_{k-1};$$

$$\tau + \tau^2 = 1;$$

$$\tau = \frac{-1 \pm \sqrt{1+4}}{2} \approx 0,61803.$$

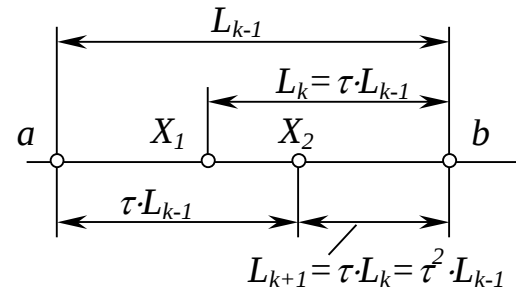


Рис. 3.3 Розбиття інтервалу локалізації

Етапи пошуку є наступними.

1. Реалізація всередині інтервалу $[a; b]$ двох експериментів з абсцисами X_1 та X_2 , зсунутими від країв на $\tau^2 \cdot L_{k-1}$.
2. Порівняння отриманих значень ЦФ. Можливі дві ситуації (рис. 3.4).
3. Звуження інтервалу локалізації шляхом відкидання однієї з крайніх частин.
4. Реалізація одного експерименту всередині інтервалу локалізації на відстані $\tau^2 \cdot L_{k-1}$ від того краю, з якого було відкинуто частину.
5. Повернення до п.2.

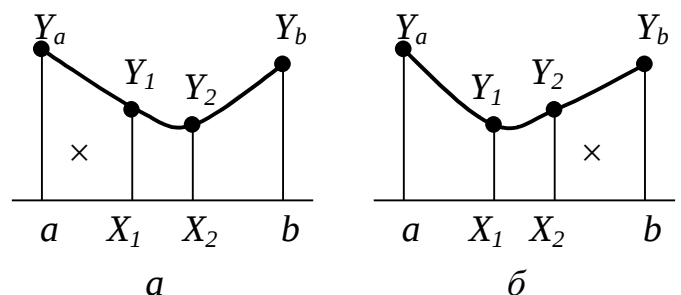


Рис. 3.4 Ситуації перед відкиданням підінтервалів

Як видно, на кожному циклі додається лише один додатковий експеримент. Це обумовлює більшу ефективність методу золотого перерізу порівняно з методом половинного ділення.

3.3 Метод квадратичної апроксимації

Метод базується на апроксимації ЦФ $Y(X)$ квадратичним поліномом за результатами трьох експериментів. Завдяки цьому враховується не лише якісне співвідношення рівнів ЦФ (більше/менше), а й величина різниці між ними (тобто неявним чином – похідна ЦФ). Це забезпечує більш швидку збіжність порівняно з уже розглянутими методами. Вимогами до ЦФ є відсутність розривів та достатня гладкість (для забезпечення стійкості пошуку). Попередня локалізація екстремуму не обов'язкова.

На рис. 3.5 суцільною лінією $Y(X)$ показано вигляд ЦФ (звичайно, наперед невідомий), штриховою $Y'(X)$ – її квадратична апроксимація, що проходить через ті ж самі три точки, що й ЦФ. Координати точок 1, 2, 3 отримано будь-яким способом на попередньому етапі.

Рівняння квадратичної параболи виходячи з цих координат

$$Y'(X) = a_0 + a_1(X - X_1) + a_2(X - X_1)(X - X_2). \quad (3.3)$$

$$\text{Для точки 1: } X = X_1; Y'(X_1) = Y(X_1) = Y_1 \text{ та } a_0 = Y_1. \quad (3.4)$$

$$\text{Для точки 2: } X = X_2; Y'(X_2) = Y(X_2) = Y_2 \text{ та } a_1 = \frac{Y_2 - Y_1}{X_2 - X_1}. \quad (3.5)$$

$$\text{Для точки 3: } X = X_3; Y'(X_3) = Y(X_3) = Y_3 \text{ та } a_2 = \frac{Y_3 - Y_1 - a_1(X_3 - X_1)}{(X_3 - X_1)(X_3 - X_2)}. \quad (3.6)$$

Для оцінки очікуваного положення екстремуму прирівнюємо похідну параболи нулю:

$$\frac{dY'}{dX} = a_1 + a_2[(X - X_1) + (X - X_2)] = 0,$$

звідки абсциса очікуваної мінімальної точки 4'

$$X_4 = \frac{(X_2 + X_1) \cdot a_2 - a_1}{2a_2}. \quad (3.7)$$

Звичайно, реальна ЦФ, на відміну від її апроксимації, екстремуму в цій точці не матиме, а результатом експерименту в ній буде значення $Y(X_4) = Y_4$. Після порівняння значень ЦФ у чотирьох точках (як на рис. 3.2) один з підінтервалів відкидається (для ситуації, зображеної на рис. 3.5, буде виключеним підінтервал $[X_1; X_2]$), а три інші точки будуть використано для отримання нової апроксимації за виразами (3.4) – (3.6) та нового прогнозу за виразом (3.7).

Модифікацією даного методу є метод Девідона-Свенна-Кемпі (ДСК). Він передбачає використання евристичного методу локалізації (рис. 3.1) з подвоєнням кроку, доки екстремум не буде взятий у вилку (точки X_1, X_2, X_3 на рис. 3.6). Після цього виконується крок назад (т. X_4), і з трьох отриманих

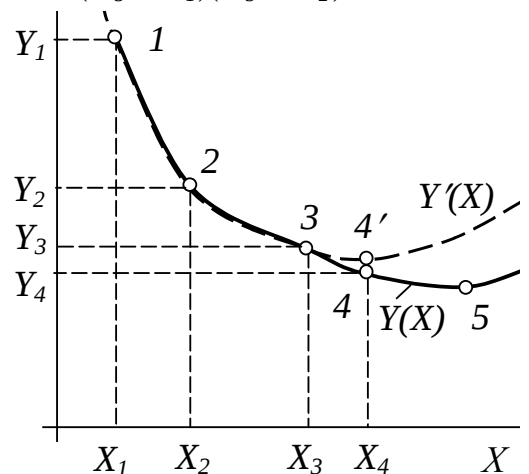


Рис. 3.5 До методу квадратичної апроксимації

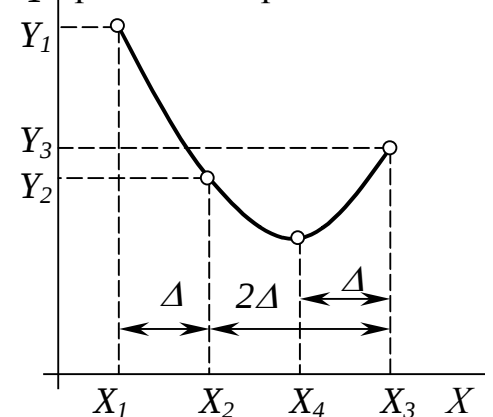


Рис. 3.6 До методу ДСК

однакових частин відкидається одна (як у методі золотого перерізу). Три точки, що лишилися, використовують для подальшої квадратичної апроксимації.

3.4 Метод Ньютона

Якщо відома похідна цільової функції $\varphi(X) = dY(X)/dX$, пошук екстремуму $Y(X)$ можна замінити на пошук нуля $\varphi(X)$. Двох точок (1 та 2 на рис. 3.7) залежності $\varphi(X)$ достатньо для першого прогнозу положення нуля. Прямая лінія 1-2 близька до дотичної лінії $\varphi(X)$ та перетинає вісь абсцис у точці $X = X_m$. З двох подібних трикутників матимемо співвідношення їх катетів:

$$\frac{X_m - X_1}{X_2 - X_1} = \frac{\varphi_1}{\varphi_2 - \varphi_1}.$$

Звідси матимемо прогноз абсциси нуля

$$X_m = \frac{\varphi_1}{\varphi_2 - \varphi_1}(X_2 - X_1) + X_1, \quad (3.8)$$

за яким отримуємо нове значення похідної φ_3 .

Якщо знаки похідної в тт.1 та 2 не відрізняється, на наступному етапі інтервал $[X_1; X_2]$ відкидають та здійснюють переприсвоєння $X_1 = X_2$ $X_2 = X_m$. Виходячи з двох останніх точок здійснюють новий прогноз положення нуля.

Рано чи пізно похідна в т.3 змінить знак і замість ситуації рис. 3.8,а матимемо ситуацію рис. 3.8,б, після чого мінімум можна вважати локалізованим.

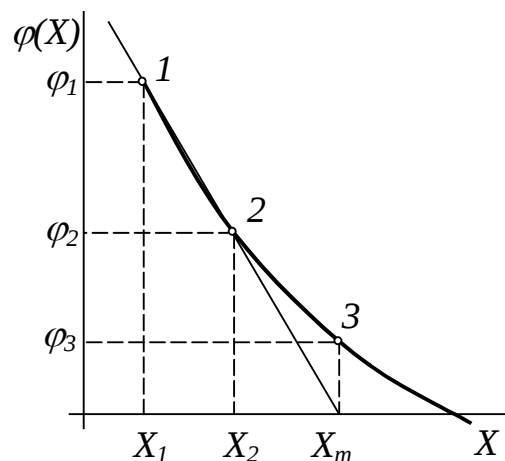


Рис. 3.7 Метод Ньютона

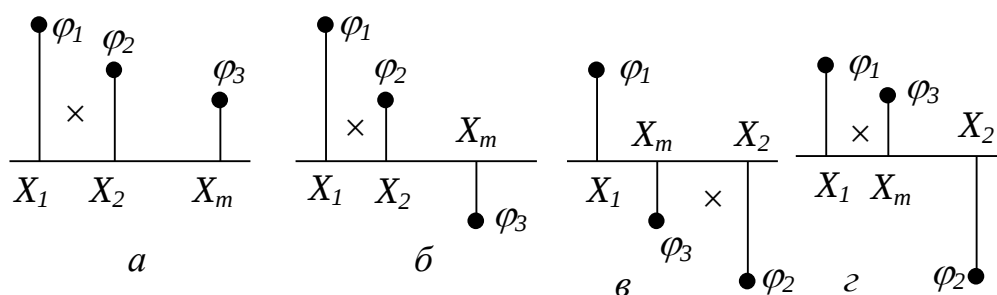


Рис. 3.8 Ситуації перед відкиданням інтервалів

На наступних етапах звуження інтервалу $\text{sign } \varphi_1 \neq \text{sign } \varphi_2$, проте знак φ_3 може бути різним. Тому у ситуації рис. 3.8,в відкидають інтервал $[X_m; X_2]$, у ситуації рис. 3.8,г – інтервал $[X_1; X_m]$. Інтервали, що відкидаються, позначені на рис. 3.8 хрестиками.

Пошук закінчують, коли нове значення $\varphi(X)$ несуттєво відрізняється від нуля або нова абсциса – від попередньої.

4 БАГАТОВИМІРНІ МЕТОДИ ПОШУКУ

Багатовимірні методи застосовують у випадках, коли ЦФ є функцією кількох факторів. Одним з критеріїв закінчення пошуку, як і для однофакторних задач, може бути умова незначної зміни ЦФ (3.1).

Що стосується ширини інтервалу локалізації, то умову (3.2) слід перевіряти для кожного з N факторів:

$$2 \left| \frac{b_j - a}{b_j + a_j} \right| \leq \delta_{x_j}. \quad (4.1)$$

Тут, як і раніше a_j та b_j – координати границь інтервалу локалізації, а δ_{x_j} – припустима ширина цього інтервалу для кожного з факторів.

4.1 Метод покоординатного пошуку Гауса-Зайделя

Хід пошуку проілюстрований рис. 4.1. Даний метод є найпростішим і передбачає почергову зміну факторів. Так, від початкової точки A_0 з координатами X_{10}, X_{20} спочатку змінюється лише координата X_1 до отримання мінімуму цільової функції (т. A_1). Після цього фактор X_1 підтримується незмінним, а змінюється лише X_2 до отримання нового мінімуму ЦФ (т. A_2). Далі процес зміни факторів повторюється. Кожна нова ітерація наближає до кінцевої мети (екстремуму в т. A). Як видно пошук у двовірному просторі складається з послідовності одновимірних пошуків. На кожному з етапів одновимірного пошуку можна використовувати будь-який з одновимірних методів, розглянутих у розділі 3.

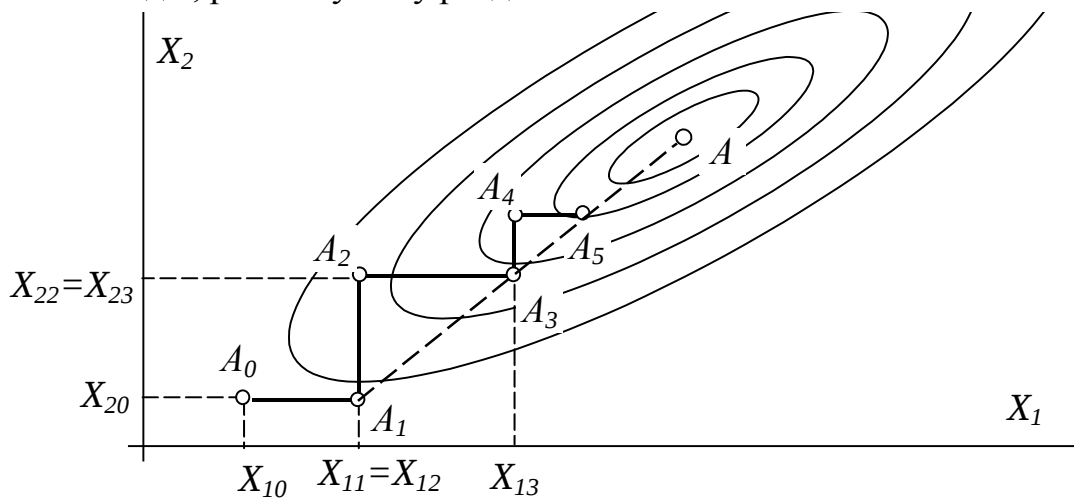


Рис. 4.1 Метод Гауса-Зайделя

До недоліків методу слід віднести невисоку швидкодію (особливо у вузькому яру, поздовжня вісь якого непаралельна осям координат, як на рис. 4.1).

4.2 Метод спряжених напрямів Пауела

Цей прямий метод базується на попередньому. З початкової точки A_0 здійснюють один, а згодом – ще N одновимірних пошуків за методом Гауса-Зайделя (для двофакторної задачі – два пошуки). Координати першого A_1 та останнього A_{N+1} мінімумів (на рис. 4.1 – тт. A_1 та A_3) визначають напрямок подальшого руху (спряжений напрям). Надалі він відбувається від т. A_3 вздовж прямої, яка з'єднує її та т. A_1 (пунктир на рис. 4.1). Координати нових точок на спряженому напрямі

$$X_j = X_{j,N+1} + h(X_{j,N+1} - X_{j,1}), \quad (4.2)$$

або для двофакторної задачі, як на рис. 4.1:

$$\begin{aligned} X_1 &= X_{1,3} + h(X_{1,3} - X_{1,1}); \\ X_2 &= X_{2,3} + h(X_{2,3} - X_{2,1}). \end{aligned} \quad (4.3)$$

Тут і надалі перший індекс у координати визначає номер фактору, другий – номер точки у факторному просторі, а N дорівнює розмірності факторного простору (тобто кількості факторів). Системи алгебраїчних рівнянь (4.2) або (4.3) пов'язують усі N факторів один з одним через додаткову змінну h . Тому фактори перестають бути незалежними, і фактично пошук уздовж спряженого напрямку перетворюється на пошук такого значення h , за якого на цьому напрямі ЦФ стає мінімальною. Геометричний сенс змінної h – відстань уздовж лінії A_1A_3 від т. A_3 , виражена у частках від гіпотенузи A_1A_3 .

Від екстремуму, що буде знайдений на спряженому напрямі, знову здійснюють N одновимірних покоординатних пошуків, за результатами яких визначають новий спряжений напрям.

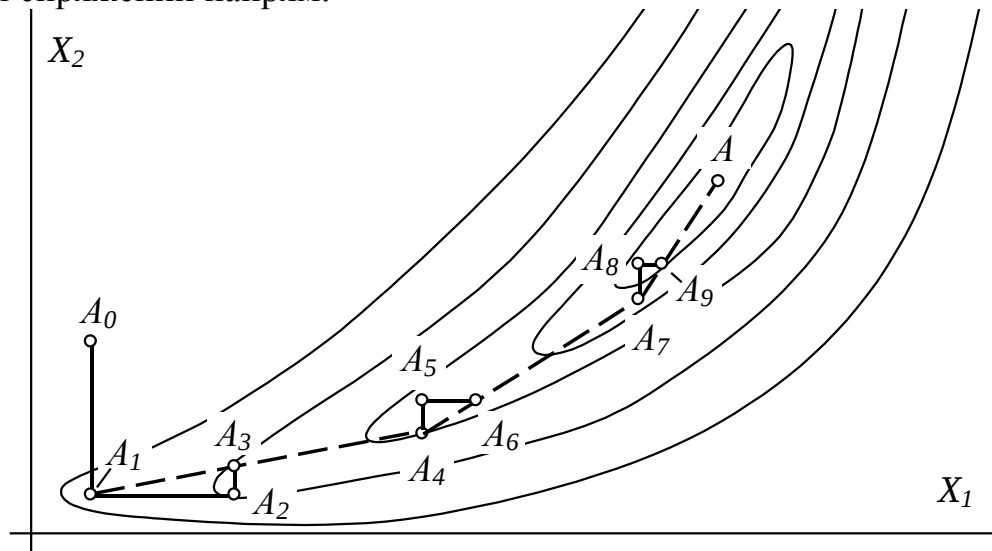


Рис. 4.2 Поведінка методу Пауела у криволінійному яру

Метод Пауела є одним з найкращих прямих методів. Якщо ЦФ є квадратичною функцією усіх факторів, екстремум буде знайдено на першому ж спряженому напрямі. Він непогано поводить у криволінійному яру (рис. 4.2), обираючи спряжений напрям уздовж поздовжньої осі яру.

4.3 Метод Хука-Дживса (пошук за зразком)

Процес пошуку складається із двох почережних етапів:

- дослідницький пошук (послідовність пробних кроків уздовж кожної з осей координат);
- пошук за зразком (один крок за всіма факторами одночасно з урахуванням результатів дослідницького пошуку).

Метою першого етапу є визначити напрям руху до мінімуму вздовж кожної з осей. З початкової (базисної) точки A_0 здійснюють пробні кроки вздовж однієї з осей

$$X_{j,i} = X_{j,0} \pm h_j,$$

доки не буде отримано краще значення ЦФ порівняно з базисною точкою. Отримана краща точка призначається новою базисною, і рухи вздовж наступних осей відбуваються вже від неї. Якщо пробні кроки вздовж певної осі не покращують ЦФ, базисна точка не змінюється. У разі невдалих дослідницьких кроків уздовж усіх осей перевіряють умову (4.1) і, якщо вона не виконується, зменшують величину кроків ($h_j = h_j/2$) до отримання вдалого результату або виконання умови локалізації.

На другому етапі остання точка вдалого дослідницького пошуку (A_N) приймається як нова базисна. Подібно до методу Пауела з неї виконують один крок уздовж прямої, яка з'єднує точки A_0 та A_N (тобто з одночасною зміною усіх факторів):

$$X_{j,N+1} = X_{j,N} + \gamma(X_{j,N} - X_{j,0}),$$

де $\gamma=1 \dots 2$.

В околицях отриманої точки здійснюють новий дослідницький пошук для визначення нового напрямку руху за зразком.

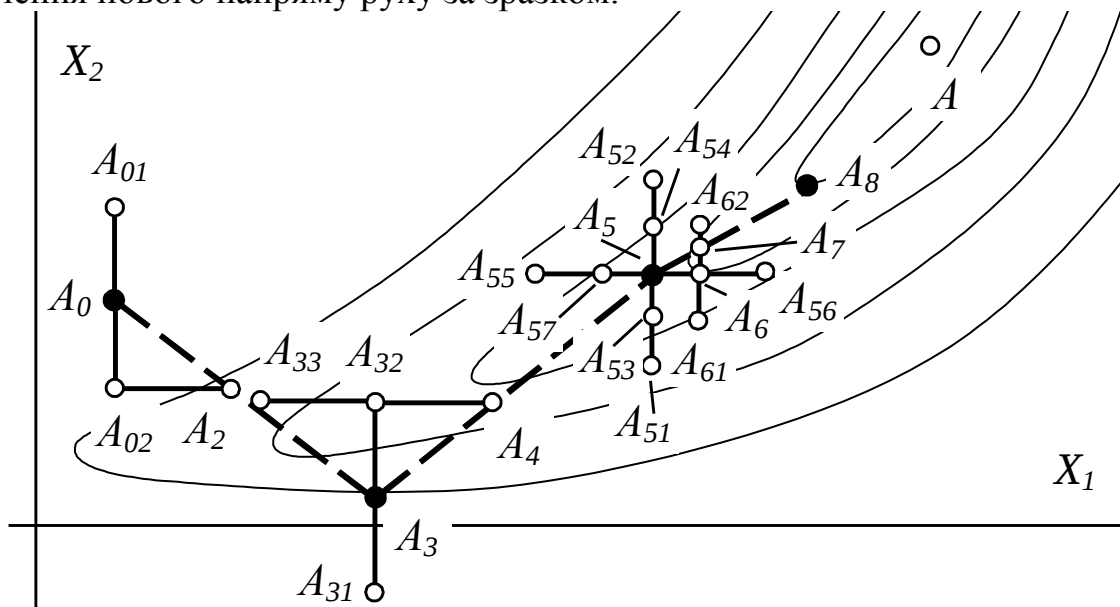


Рис. 4.3 Пошук за зразком

Якщо пошук за зразком був невдалим, повертаються до базисної точки та здійснюють новий дослідницький пошук із меншим кроком h .

На рис. 4.3 показаний процес пошуку за методом Хука-Дживса. Чорною заливкою позначено базисні точки, суцільними лініями – кроки дослідницького пошуку, штриховими – пошук за зразком. З початкової точки A_0 уздовж осі X_2 здійснено крок до т. A_{01} (невдалий), згодом – до т. A_{02} . оскільки остання виявилася кращою за т. A_0 , то кроки вздовж осі X_1 відбуваються з кращої точки (A_{02}) до т. A_2 .

Перший дослідницький пошук виявився вдалим, оскільки дав зменшення ЦФ. Тому вздовж прямої, що з'єднує тт. A_0 та A_2 , відбувається пошук за зразком з одночасною зміною обох факторів до т. A_3 . Після невдалої спроби у т. A_{31} отримано кращу точку A_{32} , потім після невдалої т. A_{32} – кращу точку A_4 . Координати точок A_3 та A_4 використовуються для нового пошуку за зразком, результатом яко-

го буде т. A_5 . Згадана точка розташована на дні яру, тому спроби в тт. A_{51}, A_{52} були невдалими. Навіть зменшення кроку (тт. A_{53}, A_{54}) не дало бажаного ефекту, тому переміщення вздовж осі X_2 припинено. Надалі дослідницький пошук відбувається вздовж осі X_1 . Точки A_{55}, A_{56} були невдалими, а після зменшення кроку – так само і т. A_{57} . Кращий результат дав лише крок до т. A_6 , з якої послідовні кроки до тт. A_{61} та A_{62} також не мали успіху. Лише після зменшення кроку отримано т. A_7 , від якої відбувся пошук за зразком до т. A_8 .

Як видно з рис. 4.3, метод є достатньо ефективним у криволінійних ярах завдяки зменшенню кроку на дні яру та скеруванню пошуку за зразком уздовж його осі.

4.4 Симплексний метод

Метод належить до категорії прямих. Базовими поняттями методу є поняття симплексу та відбиття. Симплексом називають правильну фігуру з мінімально можливою кількістю граней та однаковою довжиною ребер (на площині це рівносторонній трикутник, у тривимірному просторі – тетраедр). Кількість вершин симплексу завжди на одиницю більше розмірності факторного простору. У процесі пошуку симплекс переміщується шляхом відбиття однієї зі своїх вершин (тієї, що має найгірше значення ЦФ) відносно центру ваги інших вершин.

Першим етапом пошуку є формування початкового симплексу. Нехай координати початкової точки $A_0 \in [X_{10}, X_{20}, X_{30}, \dots, X_{j0}, \dots, X_{N0}]$. Тоді координати інших N вершин будуть:

$$X_{ji} = \begin{cases} X_{j0} + \delta_1 & \text{якщо } j \neq i; \\ X_{j0} + \delta_2 & \text{якщо } j = i, \end{cases} \quad (4.4)$$

$$\text{де } \delta_1 = \frac{\sqrt{N+1} + N - 1}{N\sqrt{2}} \cdot \alpha; \quad \delta_2 = \frac{\sqrt{N+1} - 1}{N\sqrt{2}} \cdot \alpha < \delta_1;$$

$j = 1, 2, 3, \dots, N$ – номер фактору; $i = 0, 1, 2, 3, \dots, N$ – номер вершини симплексу; α – параметр, який задає розміри симплексу (може зменшуватися на заключних стадіях пошуку з метою забезпечення заданої точності локалізації екстремуму); N – кількість факторів (розмірність факторного простору).

Приклад генерації симплексу для двовимірної задачі з вершинами у тт. A_0, A_1, A_2 зображений на рис. 4.4.

Координати т. A_1 :

$$X_{11} = X_{10} + \delta_2 = X_{10} + \frac{\sqrt{2+1}-1}{2\sqrt{2}} \cdot \alpha = X_{10} + \frac{\sqrt{3}-1}{2\sqrt{2}} \cdot \alpha;$$

$$X_{21} = X_{10} + \delta_1 = X_{10} + \frac{\sqrt{2+1}+2-1}{2\sqrt{2}} \cdot \alpha = X_{10} + \frac{\sqrt{3}+1}{2\sqrt{2}} \cdot \alpha.$$

Координати т. A_2 :

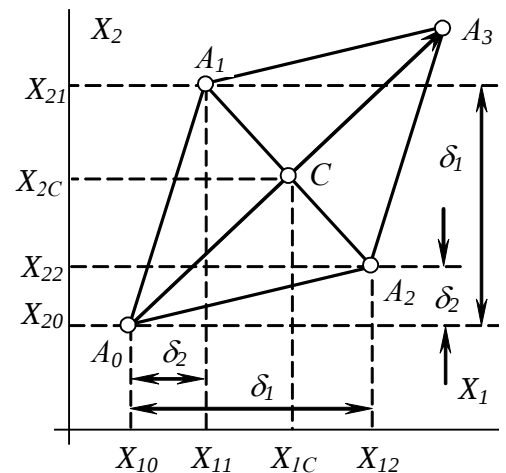


Рис. 4.4 Генерація симплексу та відбиття найгіршої вершини

$$X_{12} = X_{10} + \delta_1 = X_{10} + \frac{\sqrt{2+1}+2-1}{2\sqrt{2}} \cdot \alpha = X_{10} + \frac{\sqrt{3}+1}{2\sqrt{2}} \cdot \alpha;$$

$$X_{22} = X_{10} + \delta_2 = X_{10} + \frac{\sqrt{2+1}-1}{2\sqrt{2}} \cdot \alpha = X_{10} + \frac{\sqrt{3}-1}{2\sqrt{2}} \cdot \alpha.$$

У вершинах симплексу визначаються значення ЦФ та виявляється вершина з найгіршим її значенням.

Другим етапом пошуку є відбиття найгіршої вершини. Нехай нею буде т. A_0 на рис. 4.4. Найгірша точка вилучається з симплексу, а замість неї надалі використовується її відбиття відносно центру ваги тих точок, що лишилися. На рис. 4.4 таким центром ваги є т. C , а т. A_3 – відбиттям найгіршої т. A_0 . У загальному випадку координатами центра ваги C та нової точки-відбиття A_3 є:

$$X_{jC} = \frac{1}{N} \sum_{\substack{i=0 \\ j=0 \\ i \neq b}}^N X_{ji};$$

$$X_{jH} = X_{jC} + (X_{jC} - X_{jb}).$$

Тут індекс “ b ” має відношення до найгіршої точки, а “ H ” – до нової (відбитої). Для ситуації, зображеної на рис. 4.4:

$$X_{1C} = \frac{1}{2}(X_{11} + X_{12});$$

$$X_{2C} = \frac{1}{2}(X_{21} + X_{22});$$

$$X_{1H} = X_{1C} + (X_{1C} - X_{10});$$

$$X_{2H} = X_{2C} + (X_{2C} - X_{20}).$$

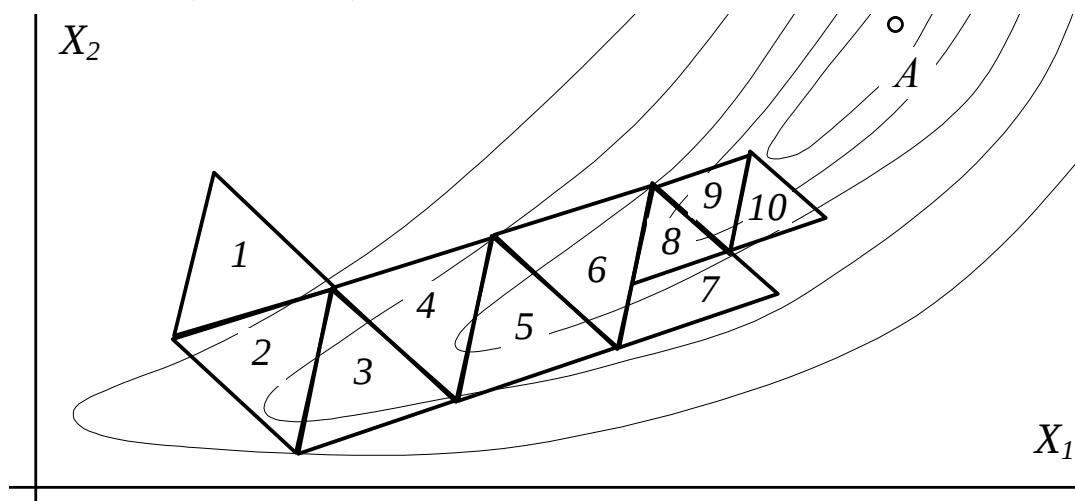


Рис. 4.5 Процес пошуку за симплексним методом

Приклад перебігу пошуку показаний на рис. 4.5. Нумерація симплексів відповідає послідовності їх формування. Видно, що після отримання симплексу 7 алгоритм зациклюється (починається взаємне відбиття симплексів 6 та 7). Для продовження пошуку дві гірші точки останнього симплексу відкидаються, а навколо точки з найкращим значенням ЦФ (т. В) формується новий симплекс 8 шляхом зменшення α у виразах (4.4). З рис. 4.5 видно, що подібне зациклювання

До переваг методу слід віднести простоту та ефективність навіть за великих похибок в обчисленні ЦФ. Недоліками є повільна збіжність (особливо у вузьких ярах).

Метод є подальшим розвитком симплексного методу. Відмінністю є можливість не лише зменшення розмірів багатогранника в процесі пошуку, а й зміни його форми.

На етапі відбиття спочатку здійснюють пробне відбиття (як у симплексному методі), а подальші дії залежать від його результатів. Координати вершини нового багатогранника визначаються як:

Значення параметру Θ :

- $\Theta = 2$, якщо результат пробного відбиття кращий за найкращий до відбиття (відбиття з розтягуванням багатогранника);
- $\Theta = 0,5$, якщо результат пробного відбиття кращий тільки за найгірший до відбиття (відбиття зі стисканням);
- $\Theta = -0,5$, якщо результат пробного відбиття гірший за найгірший до відбиття (стискання без відбиття);
- $\Theta = 1$ у всіх інших випадках.



25

($\Theta = 0,5$). Після пробного відбиття найгіршої вершини отриманого трикутника маємо точку 3, яка є кращою за найкращу. Тому замість неї надалі використовується т. 4, що розташована вдвічі далі від центру ваги ($\Theta = 2$). На наступному кроці з'являється т. 5 ($\Theta = 1$). Далі маємо пробну точку 7, кращу лише за найгіршу. Тому замість неї вводиться т. 7 ($\Theta = 0,5$). Замість пробної т. 8 маємо т. 9, замість т. 10 – т. 11, замість т. 12 – т. 13, розташовану досить близько до шуканого екстремуму.

Розглянутий метод є достатньо ефективним в задачах безумовної мінімізації, якщо кількість факторів не перевищує 6 [9].

4.6 Метод комплексів Бокса

Метод чимось подібний до методу Нелдера-Міда. Пошук так само складається з двох етапів: генерація початкової сукупності точок (комплексу) та подальше відбиття найгіршої з них. Метод може бути використаний у задачах з обмеженнями.

Процес генерації початкової сукупності точок (комплексу) для задачі з N факторами полягає у наступному (рис. 4.7). За допомогою генератора випадкових чисел отримують вектор довжиною N , який задає координати наступної точки сукупності. Якщо перша точка потрапляє до забороненої області, вона відкидається, доки не буде отримана точка у допустимій області (на рис. 4.7 це т. 1). Якщо після цього наступна точка також опиняється у забороненій області (т. 2), її зсувають на половину відстані до центру ваги вже згенерованих точок у допустимій області, доки нова точка стане допустимою (т. 3). На рис. 4.7 стрілками показаний напрям зсуву або відбиття найгірших точок. Сумарна кількість точок не повинна бути меншою від $N+1$ (в [3] рекомендовано $\approx 2N$). На рис. 4.7 останньою (третьою) точкою сукупності є т. 4.

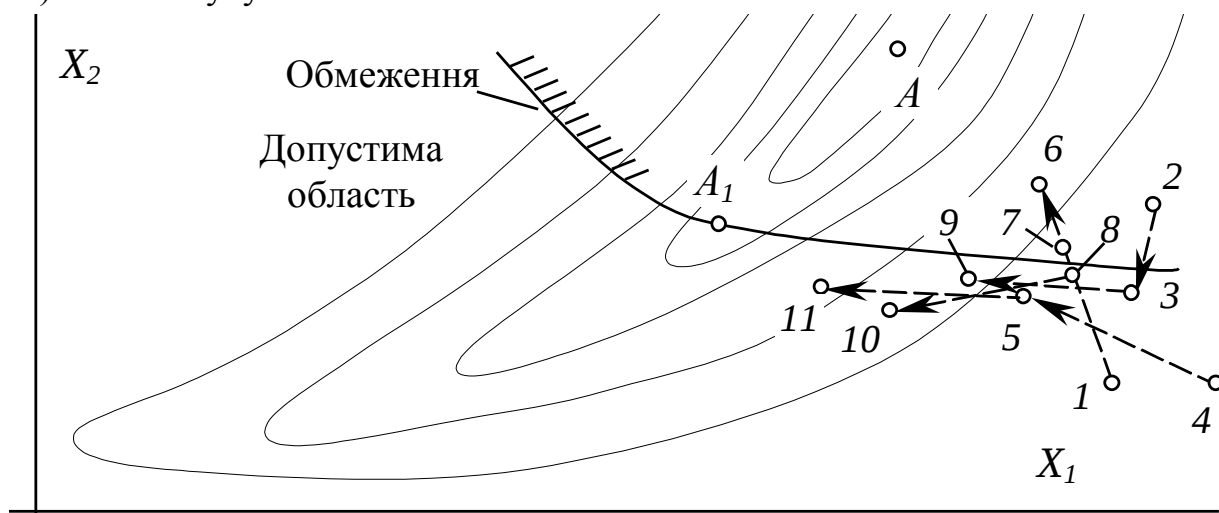


Рис. 4.7 Пошук за методом комплексів Бокса

Очевидно, що найгіршою з точок 1, 3, 4 є остання. Вона й відбивається відносно центру ваги двох інших (нова точка – т. 5). З трьох точок нового комплексу (1, 3, 5) відбиттю підлягає т. 1. Її відбиттям є т. 6 у забороненій області. Зсув останньої точки вбік середини центру ваги двох інших дає бажаний результат лише після другої спроби (тт. 7 та 8). У комплексі 3, 5, 8 найгіршою є т. 3. Її відбиттям є

т. 9. На наступних кроках формуються комплекси 5, 9, 10 та 9, 10, 11. Як видно, у процесі пошуку комплекс переміщується вздовж лінії обмеження у напрямку до точки умовного мінімуму A_1 .

Координати нової точки розраховуються за тією ж формулою, що й за попереднім методом:

$$X_{jH} = X_{jC} + \alpha \cdot (X_{jC} - X_{jb}).$$

У ній параметр α визначають таким чином:

- початкове значення $\alpha=1,3$;
- відбита точка є допустимою та не найгіршою (α не змінюється, точка залишається);
- нова точка потрапила до забороненої області (α зменшують удвічі стільки разів, скільки необхідно до потрапляння до допустимої області);
- нова точка найгірша (α зменшують удвічі стільки разів, скільки необхідно для того, щоб найгіршою стала інша точка).

4.7 Випадковий пошук

Розглянемо два варіанти випадкового пошуку: пошук з поверненням та пошук за найкращою пробою. Обидва варіанти складаються з двох етапів: етапу “обмацування” околиці базової точки та етапу вибору нової базової точки. Обидва варіанти передбачають, що навкруги базової точки (т. 0 на рис. 4.8) виконують пробні кроки фіксованої величини r у випадкових напрямках.

Особливістю першого варіанту є відсутність обмеження на кількість таких пробних кроків. Якщо значення ЦФ у новій точці гірше за її значення в базовій точці, здійснюють повернення до базової точки (звідки й назва методу) для виконання наступного пробного кроку. У разі покращання значення ЦФ етап обмацування припиняється, а останню пробну точку беруть як нову базову (т. 4 на рис. 4.8).

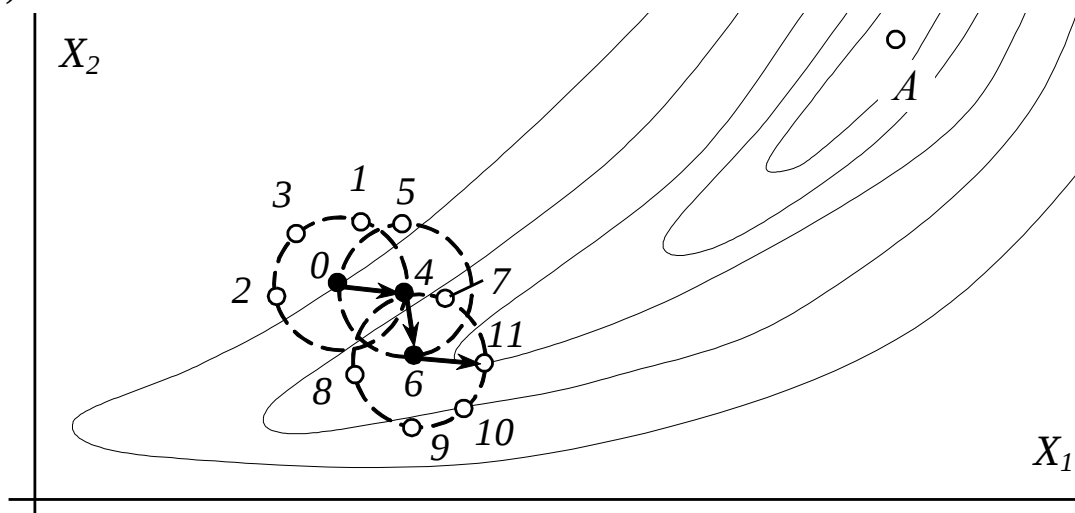


Рис. 4.8 Процес випадкового пошуку з поверненням

Навкруги цієї базової точки виконують певну кількість кроків (тт. 5, 6), доки не буде отримане краще значення ЦФ (т. 6 прийнята за наступну базову). Далі новою базовою точкою стала т. 11. З рис. 4.8 видно, що в околицях базової т. 0 здійснено чотири пробних кроки, навколо т. 4 – два, навколо т. 6 – п’ять. Оскільки

ки довжина кроків з базової точки незмінна, усі пробні точки лежать на колі з центром у базовій точці та радіусом r (на рис. 4.8 відповідні кола показані штриховими лініями, а стрілками позначений напрям переміщення базової точки). Базові точки на рис. 4.8 зашиті чорним.

Пошук за найкращою пробою відрізняється від попереднього лише одним: наперед заданою кількістю пробних кроків (у прикладі на рис. 4.9 їх п'ять). Після виконання заданої кількості пробних кроків найкраща (не обов'язково остання) з отриманих точок стає новою базовою. Якщо кількість пробних кроків достатньо велика, траєкторія руху наближається до найкращої (тобто такої, якою потече вода з початкової базової точки до дна западини).

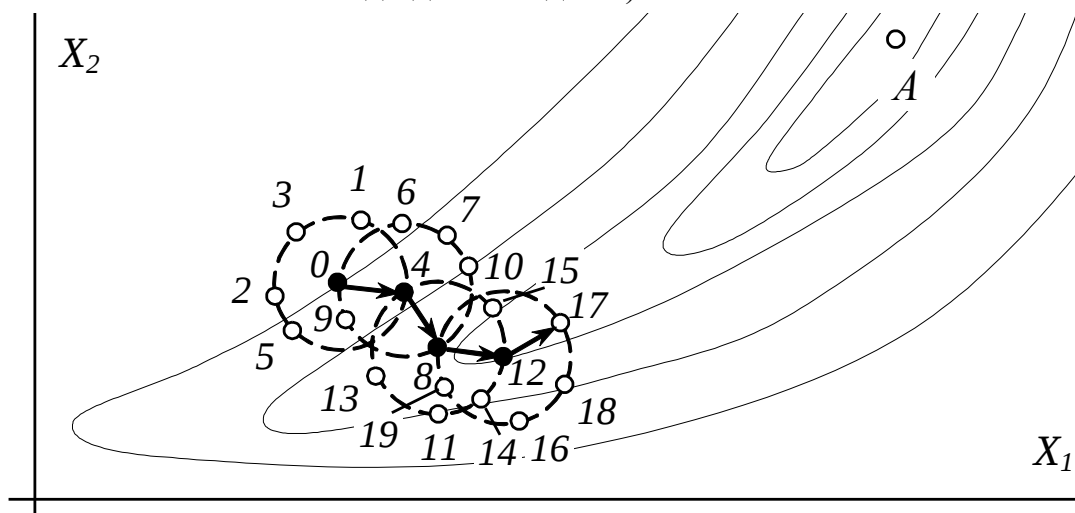


Рис. 4.9 Процес випадкового пошуку за найкращою пробою

Точність визначення положення екстремуму визначається радіусом кола r , тому на останніх етапах пошуку, коли покращення ЦФ практично не спостерігається, радіус слід зменшувати.

4.8 Метод найшвидшого спуску

Інші назви методу: метод крутого сходження, метод Бокса-Уілсона (за прізвищами авторів), метод Коші (іноді). Теоретичною основою методу є теорія планування експерименту (ТПЕ) [4, 5]. Перш ніж говорити про алгоритм пошуку екстремуму за цим методом, розглянемо деякі основні положення теорії планування багатofакторних експериментів, необхідні для розуміння подальшого матеріалу.

Однією з основних проблем, що вирішуються за допомогою ТПЕ, є дослідження поведінки ЦФ в певних (достатньо невеликих) околицях заданої точки факторного простору. Кінцевим результатом такого дослідження звичайно є алгебраїчне рівняння (поліном), який пов'язує ЦФ з незалежними змінними. Таке рівняння часто називають рівнянням регресії і використовують для прогнозу (інтерполяції або екстраполяції). Точність прогнозу залежить від порядку поліному. Теорія планування експерименту дає можливість таким чином розташувати експериментальні точки, щоб кількість їх була мінімальною, алгоритм розрахунку коефіцієнтів рівняння регресії – найпростішим, а саме рівняння – найбільш адекватним.

Для реалізації методу найшвидшого спуску достатньо мати лінійну апроксимацію поверхні відгуку. Очевидно, що лінійну апроксимацію однофакторної функції $Y = f(X)$ можна отримати за результатами експериментів лише у двох точках (тобто єдиний фактор матиме два рівні). На рис. 4.10 крива $Y(X)$ відображає ЦФ, пряма $\hat{Y}(X)$ – її лінійну апроксимацію. Пряма проведена через точки з абсцисами $X = X_{\min}$ та $X = X_{\max}$, які відстоять від точки $X = X_0$ на однакову відстань ΔX (інтервал варіювання). Очевидно, що зі зменшенням інтервалу ΔX пряма $\hat{Y}(X)$ наближатиметься до дотичної до кривої $\hat{Y}(X)$ у точці $X = X_0$.

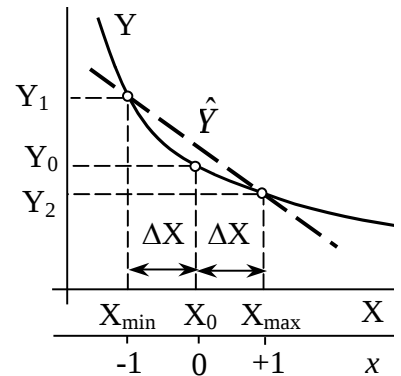


Рис. 4.10 Лінійна апроксимація функції однієї змінної

Для спрощення розрахунків замість іменованих значень фактору X використовують безрозмірні (відносні):

$$x = \frac{X - X_0}{\Delta X}. \quad (4.5)$$

Завдяки цьому новий безрозмірний фактор завжди має значення $x = -1$ або $x = +1$ (рис. 4.10).

Якщо ЦФ $Y = f(X_1, X_2)$ є функцією двох факторів, для отримання її лінійної апроксимації в околицях точки $[X_{10}, X_{20}]$ необхідно реалізувати експерименти в чотирьох точках, які утворюють квадрат у факторному просторі (рис. 4.11). Сукупність таких точок називають планом експерименту.

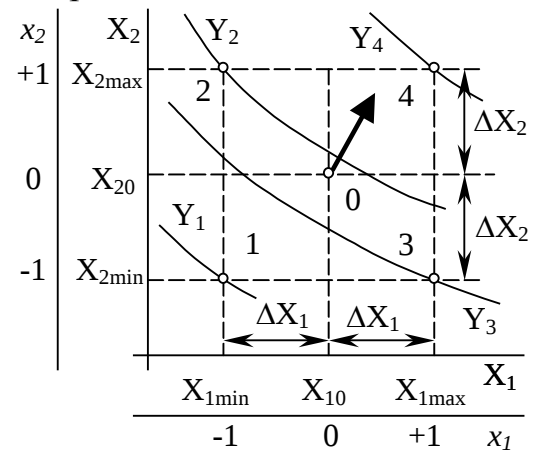


Рис. 4.11 План двофакторного експерименту

Оскільки згадані точки відображають усі припустимі сполучення значень факторів, такий план називають планом повного факторного експерименту (ПФЕ). Кількість точок ПФЕ завжди дорівнює $n = 2^N$ (для двофакторного ПФЕ $n = 2^2 = 4$). Рядки з номерами 1-4 у табл. 4.1 відповідають точкам з відповідними номерами на рис. 4.11. Для кожної точки ПФЕ після реалізації експериментів визначаються значення ЦФ (стовпчик Y у табл. 4.1), за якими розраховуються коефіцієнти

$$b_0 = \frac{\sum_{i=1}^n Y_i}{n};$$

$$b_1 = \frac{\sum_{i=1}^n x_{1i} Y_i}{n};$$

$$b_2 = \frac{\sum_{i=1}^n x_{2i} Y_i}{n}$$

поліному $\hat{Y} = b_0 + b_1 x_1 + b_2 x_2$, який є ліній-

Таблиця 4.1

План двофакторного експерименту

№	X_1	x_1	X_2	x_2	Y	$\hat{Y}$
1	$X_{1\min}$	-1	$X_{2\min}$	-1	Y_1	$\hat{Y}_1$
2	$X_{1\min}$	-1	$X_{2\max}$	+1	Y_2	$\hat{Y}_2$
3	$X_{1\max}$	+1	$X_{2\min}$	-1	Y_3	$\hat{Y}_3$
4	$X_{1\max}$	+1	$X_{2\max}$	+1	Y_4	$\hat{Y}_4$
0	X_{10}	0	X_{20}	0	Y_0	$\hat{Y}_0$

ною апроксимацією ЦФ. За достатньо малого значення інтервалів варіювання факторів цей поліном описує площину, дотичну до поверхні $Y = f(X_1, X_2)$ в точці $[X_{10}, X_{20}]$.

У загальному випадку для довільної кількості факторів поліном (т. зв. рівняння регресії) та його коефіцієнти мають вигляд

$$\hat{Y} = b_0 + \sum_{j=1}^N b_j x_j; \quad b_0 = \frac{\sum_{i=1}^n Y_i}{n}; \quad b_j = \frac{\sum_{i=1}^n x_{ji} Y_i}{n}, \quad (4.6)$$

де $i = 1 \dots n$ – номер рядка плану ПФЕ; $j = 1 \dots N$ – номер фактора.

Якщо кількість факторів $n = 3$, точки плану розташовані у вершинах кубу, за більшої кількості факторів – гіперкубу. Коефіцієнти b_j є частинними похідними ЦФ за відповідними факторами. Оскільки у кожному рядку плану прогнозоване значення ЦФ завжди дорівнює реальному її значенню ($\hat{Y}_i = Y_i$), для оцінки адекватності отриманого поліному необхідно поставити додатковий експеримент у центрі плану в точці, де всі фактори $x_i = 0$ (останній рядок табл. 4.1 та т.0 на рис. 4.11). Якщо різниця $\hat{Y}_0 - Y_0$ менша від припустимої, поліном вважається адекватним, якщо більша, слід зменшити інтервали варіювання факторів та повторити ПФЕ.

Якщо відоме рівняння гіперплощини, дотичної до гіперповерхні відгуку, легко знайти напрям руху до екстремуму, тобто напрям градієнту. Він збігається з прямою, перпендикулярною до зазначеної гіперплощини. У двофакторному просторі ця пряма буде перпендикулярною до ліній рівня ЦФ (товста лінія зі стрілкою на рис. 4.11). Щоб рухатися вздовж цієї прямої, фактори повинні змінюватися згідно рівнянь:

$$\left. \begin{aligned} X_1 &= X_{10} + r b_1 \Delta X_1; \\ X_2 &= X_{20} + r b_2 \Delta X_2; \\ &\dots \\ X_j &= X_{j0} + r b_j \Delta X_j; \\ &\dots \\ X_N &= X_{N0} + r b_N \Delta X_N; \end{aligned} \right\} \quad (4.7)$$

У цій системі N рівнянь $N+1$ невідомих (N факторів та нова змінна r). Це означає, що r – єдина незалежна змінна, змінюючи яку та розраховуючи значення факторів за (4.7), можна знайти екстремум на згаданій прямій. У системі рівнянь (4.7) $r > 0$ для пошуку максимуму та $r < 0$ для пошуку мінімуму ЦФ.

Таким чином, алгоритм пошуку за методом найшвидшого спуску полягає у наступному:

- у найближчих околицях початкової точки (A_1 на рис. 4.12) реалізується ПФЕ, за результатами якого визначаються коефіцієнти поліному (4.6) як вихідні дані для подальшого руху;

- здійснюється рух до екстремуму вздовж прямої A_1A_2 згідно з системою (4.7) з метою пошуку такого r , що відповідає найкращому значенню ЦФ (т. A_3);
- навколо отриманої точки реалізується новий ПФЕ, щоб отримати можливість для нового одновимірного пошуку (пряма A_3A_4) тощо.

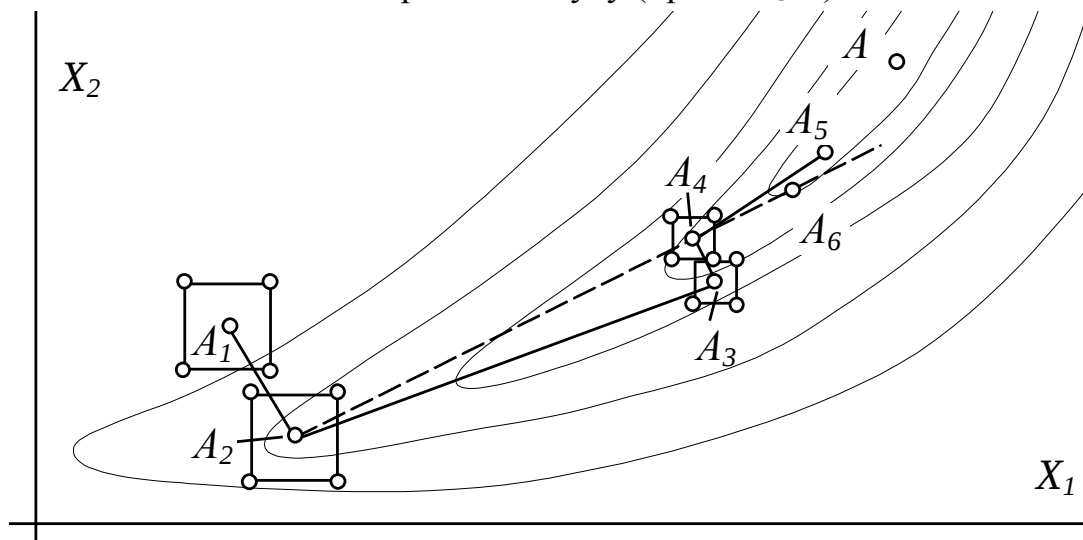


Рис. 4.12 Пошук за методом найшвидшого спуску (двофакторна задача)

Якщо наступні пошуки стають невдалими (не призводять до покращання ЦФ), інтервали варіювання факторів ΔX_j зменшують (як на рис. 4.12) продовжують пошук.

Таким чином, як і у більшості методів, відбувається чергування багатфакторних етапів (реалізація ПФЕ) та одно факторних (пошук вздовж прямої). Як і завжди, в процесі одновимірного пошуку (рух вздовж напрямку градієнту) можна використовувати будь-які одновимірні методи. Даний метод належить до непрямих (використовуються похідні ЦФ) і є одним з найкращих з точки зору швидкодії.

4.9 Метод спряжених градієнтів Флетчера-Рівса

Метод подібний до методу спряжених напрямків Пауела (п. 4.2) і відрізняється лише тим, що замість $N+1$ одновимірних пошуків за методом Гауса-Зайделя здійснюють таку саму кількість пошуків за методом найшвидшого спуску. Потім, так само як і в методі Пауела, відбувається пошук вздовж спряженого напрямку (тобто вздовж лінії, яка з'єднує кінцеві точки першого та $N+1$ -ого найшвидших спусків). На рис. 4.12 спряженим напрямком буде лінія A_2A_4 . Після отримання екстремуму на спряженому напрямі (т. A_6) знову здійснюють N пошуків за методом найшвидшого спуску для отримання нового спряженого напрямку тощо. Метод є найкращим за швидкістю.

4.10 Особливості оптимізації з обмеженнями

Загальним формулюванням задачі оптимізації з обмеженнями (умовної оптимізації) є: знайти екстремум ЦФ $Y = f(X)$ за умови, що $h(X) = 0$ та (або) $g(X) \geq 0$. Тут $h(X)$ та $g(X)$ – обмеження у вигляді відповідно рівності та нерівності.

Якщо рівняння $h(X)=0$ розв'язується відносно одного з факторів, з нього можна отримати вираз $X_j = \sigma(X_1, X_2, \dots, X_N)$. Останній можна підставити до рівняння ЦФ, врахувавши тим самим обмеження та виключивши фактор X_j з переліку незалежних змінних. Це призведе до зменшення розмірності факторного простору на одиницю. При цьому під час подальшого пошуку автоматично буде забезпечений рух уздовж лінії $h(X)=0$. Проте в технічних задачах таку ситуацію можна вважати рідкісною вдачею. У більшості випадків обмеження (як рівності, так і нерівності) є складними нелінійними функціями факторів та не розв'язуються відносно останніх. Тому виходом є лінеаризація обмеження (заміна реальної функції $h(X)=0$ або $g(X)=0$ її лінійною апроксимацією

$$a_0 + a_1 X_1 + a_2 X_2 + \dots + a_N X_N = 0 \quad (4.8)$$

в околицях поточної точки факторного простору). Площина, яка описується функцією (4.8), є дотичною до дійсної поверхні обмеження $h(X)=0$. Відомо, що поліном першого порядку завжди розв'язується відносно будь-якого з факторів. Наприклад:

$$X_1 = (-a_0 - a_2 X_2 - \dots - a_N X_N) / a_1. \quad (4.9)$$

Розглянемо принцип використання лінійних апроксимацій обмежень на прикладі методу найшвидшого спуску.

Лінеаризація обмежень у вигляді рівності

Проблемою є те, що рух уздовж дотичної одразу ж призводить до відходу від поверхні обмеження та порушення цього обмеження. Основними ідеями є:

- призначення навколо обмеження певного “коридору”, всередині якого у процесі пошуку тимчасово дозволено порушувати обмеження на величину $\pm \Delta$;
- зміна пріоритетів (якщо поточна точка лежить у межах дозволеного коридору, метою є мінімізація ЦФ, якщо за межами коридору – за мету вважається вихід на лінію обмеження $h(X)=0$);
- під час мінімізації ЦФ відбувається рух дотичною до обмеження, доки не буде знайдений мінімум усередині коридору або не відбудеться вихід на межу коридору;
- під час руху до лінії обмеження $h(X)=0$ значення ЦФ ігноруються.

На рис. 4.13 показаний процес пошуку для двовимірної задачі з обмеженням у вигляді рівності. Шуканий умовний екстремум розташований в т. A .

Початковою точкою є т. A_0 . Оскільки вона розташована у забороненій зоні, метою першого пошуку є вихід на лінію обмеження $h(X)=0$. Навколо цієї точки реалізується ПФЕ та визначаються коефіцієнти поліному

$$\hat{h} = a_0 + a_1 X_1 + a_2 X_2,$$

який апроксимує поверхню обмеження $h = h(X)$.

Рух відбувається вздовж прямої, перпендикулярної лініям рівня обмеження у напрямку до $h=0$ згідно з системою рівнянь

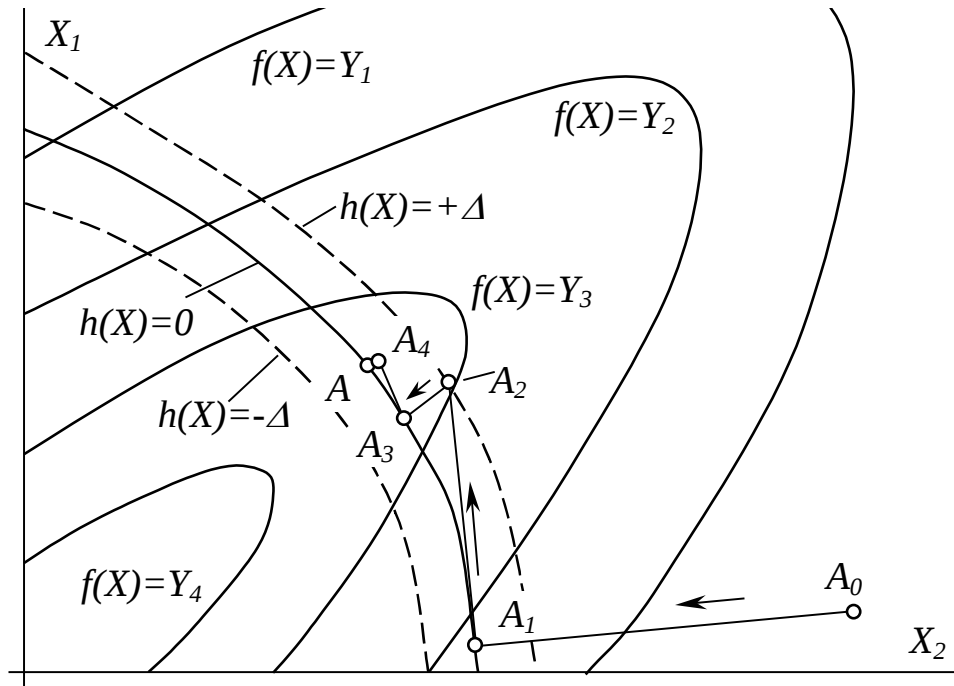


Рис. 4.13 Хід пошуку з обмеженням у вигляді рівності

$$\left. \begin{aligned} X_1 &= X_{10} + ra_1 \Delta X_1; \\ X_2 &= X_{20} + ra_2 \Delta X_2. \end{aligned} \right\} \quad (4.10)$$

У цій системі рівнянь лише одна незалежна змінна r . Тому розмірність факторного простору зменшується на одиницю, а задача перетворюється на однофакторну.

Пошук закінчується у т. A_1 на лінії обмеження. Після цього навколо т. A_1 реалізується наступний ПФЕ з метою отримання коефіцієнтів поліномів

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 \quad \text{та} \quad (4.11)$$

$$\hat{h} = a_0 + a_1 X_1 + a_2 X_2 = 0. \quad (4.12)$$

Якби факторів було більше за два, з т. A_1 відбувся би пошук у площині, заданий рівнянням (4.12) у напрямі зменшення ЦФ. Розташування поточних точок пошуку на згаданій площині забезпечувалося б розв'язанням рівняння (4.12) відносно одного з факторів та зв'язуванням цього фактора з іншими (розмірність факторного простору зменшилася би на одиницю). Одновимірний пошук мінімуму ЦФ при цьому забезпечувався би зміною r у системі рівнянь типу (4.7), з якої виключений фактор, зв'язаний рівнянням (4.12).

Для двофакторної ж задачі на рис. 4.13 рух уздовж лінії обмеження до мінімуму ЦФ забезпечується зростанням X_2 , коли X_1 пов'язаний з ним рівнянням (4.12):

$$X_1 = (-a_0 - a_2 X_2) / a_1.$$

Кінцевою точкою цього пошуку буде т. A_2 на межі коридору (рис. 4.13). Після цього змінюється пріоритет: навколо т. A_2 реалізується ПФЕ з метою отримання коефіцієнтів поліному вигляду (4.12) та наступного пошуку нуля обмеження (як в околицях т. A_0). Результатом пошуку є т. A_3 на лінії обмеження. Наступним рухом буде пошук мінімуму ЦФ уздовж дотичної до обмеження. Як видно з рис. 4.13, на цій дотичній усередині коридору буде знайдено мінімум (т. A_4) у

близьких околицях т. A . Для уточнення положення екстремуму можна звузити коридор, зменшивши Δ , та продовжити пошук з т. A_4 .

Лінеаризація обмежень у вигляді нерівності

Особливістю задачі з обмеженнями-нерівностями є те, що коридор призначається лише у забороненій зоні (ліворуч і нижче лінії $g(X)=0$ на рис. 4.14). Якщо рух починається з т. A_0 у забороненій області, перебіг його подібний до такого на рис. 4.13:

- вихід на лінію обмеження у т. A_1 ,
- рух усередині коридору уздовж обмеження у бік зменшення ЦФ до т. A_2 ,
- повернення на лінію обмеження у т. A_3 ,
- рух уздовж обмеження у бік шуканого екстремуму.

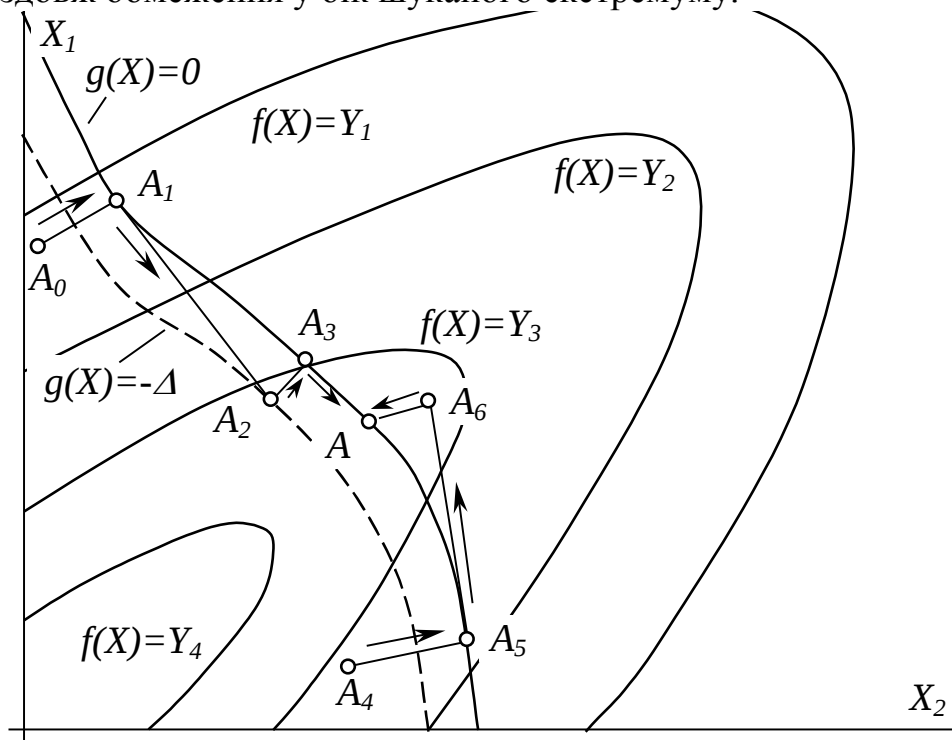


Рис. 4.14 Хід пошуку з обмеженням-нерівністю

Якщо обмеження не є опуклим (як на рис. 4.14), можлива ситуація, коли після виходу з забороненої області на лінію обмеження (з т. A_4 до т. A_5) подальший рух уздовж обмеження відбувається в допустимій області, де обмеження не є критичним (т. A_6).

4.11 Метод штрафних функцій

Метод використовується для задач з обмеженнями та базується на перетворенні задачі оптимізації. Нехай спочатку задачу сформульовано у вигляді: знайти екстремум ЦФ $Y = f(X)$ за умови, що $h(X)=0$ та $g(X) \leq 0$. Тут обмежень у вигляді як рівності $h(X)=0$, так і нерівності $g(X) \geq 0$ може бути декілька. У новому формулюванні до вихідної ЦФ $Y = f(X)$ додається штрафна функція, значення якої залежить від ступеню порушення обмежень:

$$Z = Y + S[R, h(X), g(X)],$$

де R – вага штрафу (або ціна порушення обмеження), S – штрафна функція.

Після подібного перетворення умовна задача стає безумовною і може бути розв'язана за допомогою будь-якого безумовного методу. Залежно від характеру обмежень використовують штрафні функції (ШФ) різного виду.

Симетричні (двобічні) ШФ, значення яких дорівнюють сумі квадратів або модулів порушених обмежень, придатні для задач з обмеженнями типу рівності (рис. 4.15):

- $S[h(X)] = R \sum_{i=1}^m [h_i(X)]^2$;
- $S[h(X)] = R \sum_{i=1}^m |h_i(X)|$,

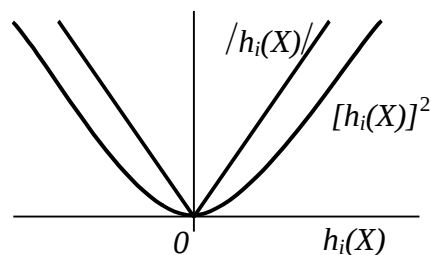


Рис. 4.15 Вигляд симетричних ШФ

де i – номер обмеження, m – їх загальна кількість.

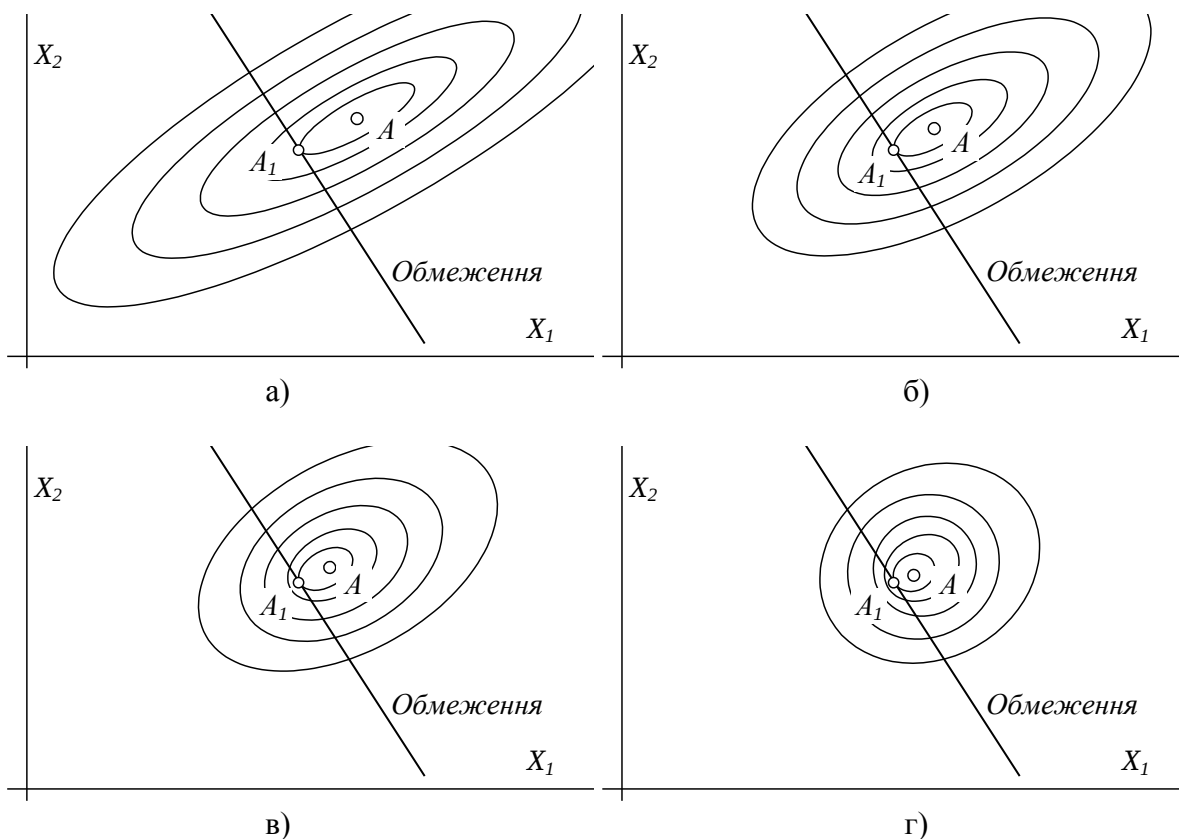


Рис. 4.16 Зміна ліній рівня двофакторної ЦФ зі збільшенням ваги штрафу R

Симетрична ШФ дорівнює нулю на лінії обмеження і зростає зі збільшенням відстані від цієї лінії (тобто штрафується порушення обмеження). Це призводить до зміни топографії поверхні відгуку: на лінії обмеження ординати нової ЦФ $Z=Y$, а на периферії поверхні $Z = f(X)$ «підіймається», причому тим сильніше, чим більша вага штрафу R . Це призводить до «сплюснення» ліній рівня ЦФ до лінії обмеження (рис. 4.16) та наближення безумовного мінімуму нової ЦФ Z (т. A) до умовного шуканого екстремуму (т. A_1). Чим більша вага штрафу R , тим сильніше «сплюснення» ліній рівня та тим ближче т. A до т. A_1 . Проте внаслідок надмірного збільшення R виникає вузький яр навколо ліній обмеження, який упові-

льнює процес пошуку. Тому задачу звичайно розв'язують у кілька етапів. Перший пошук здійснюють за невеликого значення ваги штрафу, а на кожному з наступних його збільшують у 4...5 разів. Пошук завершено, якщо наступний екстремум практично не відрізняється від попереднього.

Для обмежень-нерівностей використовують несиметричні ШФ:

- зовнішні;
- внутрішні (бар'єрні).

Внутрішні ШФ відмінні від нуля у дозволений області та необмежено зростають з наближенням до межі забороненої області (тобто виникає бар'єр, який перешкоджає виходу в заборонену область). Найбільш розповсюджені внутрішні ШФ:

- логарифмічна $S[g(X)] = -R \sum_{i=1}^k \ln[g_i(X)]$;
- гіперболічна $S[g(X)] = R \sum_{i=1}^k \frac{1}{g_i(X)}$.

Оскільки ці функції мають розрив на лінії обмеження, пошук слід починати лише з дозволених області. З тієї ж причини є небезпека виникнення автоколивань, якщо не обмежувати крок пошуку. Пошук з такими ШФ починають з достатньо великих значень ваги штрафу, яку згодом зменшують.

Зовнішні ШФ відмінні від нуля лише у забороненій області та зростають при заглибленні в неї. Звичайно використовують два види зовнішніх штрафних функцій (рис. 4.17):

- $S[g(X)] = R \sum_{i=1}^k \langle g_i(X) \rangle$;
- $S[g(X)] = R \sum_{i=1}^k \langle g_i(X) \rangle^2$,

де $\langle g_i(X) \rangle = \begin{cases} g_i(X), & \text{якщо } g_i(X) > 0 \\ 0, & \text{якщо } g_i(X) < 0 \end{cases}$ – функція зрізання.

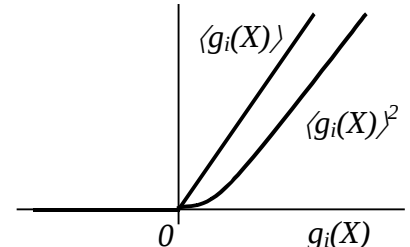


Рис. 4.17 Вигляд зовнішніх ШФ

Перевагою зовнішніх штрафів є більша збіжність, проте знайдений екстремум завжди розташований у забороненій області. Крім того, ім. віддають перевагу тоді, коли складно підібрати початкове наближення в допустимій області.

Метод ШФ нерідко використовують для отримання приблизного розв'язку, щоб з отриманої точки рухатися до екстремуму іншим, більш точним методом.

4.12 Лінійне програмування

До задач лінійного програмування (ЛП) належать ті, в яких ЦФ та обмеження є лінійними функціями факторів. Методи ЛП широко використовуються для розв'язання економічних, організаційних, промислових, військових задач. Загальним формулюванням задачі ЛП є:

$$Y = \sum_{i=1}^n a_i X_i \Rightarrow \min$$

$$\text{за умови } h_1 = \sum_{i=1}^n c_{1i} X_i \leq b_1,$$

$$h_2 = \sum_{i=1}^n c_{2i} X_i \leq b_2,$$

...

$$h_m = \sum_{i=1}^n c_{mi} X_i \leq b_m.$$

Деякі з обмежень можуть бути рівностями. Теж саме у матричній формі:

$$\mathbf{Y} = \mathbf{A} * \mathbf{X} \Rightarrow \min \text{ за умови } \mathbf{C} * \mathbf{X} \leq \mathbf{B},$$

де $\mathbf{A} = [a_1 \ a_2 \ \dots \ a_n]$ – вектор оцінок; $\mathbf{X} = [X_1 \ X_2 \ \dots \ X_n]^T$ – вектор змінних;

$\mathbf{B} = [b_1 \ b_2 \ \dots \ b_m]^T$ – вектор ресурсів;

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \dots & \dots & \dots & \dots \\ c_{m1} & c_{m2} & \dots & c_{mn} \end{bmatrix} \text{ – матриця коефіцієнтів.}$$

Розглянемо як приклад одну з типових задач ЛП.

У відділі технічного контролю (ВТК) працюють контролери III та IV розрядів. Змінна норма ВТК становить 1800 виробів. Контролер III розряду здатний перевірити 25 виробів на годину, не помиляючись у 98% випадків, контролер IV розрядів – відповідно 15 виробів та 95%. Заробітна платня контролера III розряду складає 3 €/годину, контролера IV розряду 4 €/годину. Кожна помилка контролера коштує компанії 2 €. Можна використовувати до 8 контролерів IV розряду та до 10 – третього. Яким є оптимальний склад ВТК, що забезпечує мінімум загальних витрат?

Позначимо через X_1 та X_2 кількість контролерів відповідно IV розряду та III розрядів. Загальні витрати (Y) складаються з витрат на заробітну платню (ЗП) та збитків від помилок контролерів (ЗБ):

$$\text{ЗП} = X_1 \cdot 4 \cdot 8 + X_2 \cdot 3 \cdot 8 = 32 X_1 + 24 X_2;$$

$$\text{ЗБ} = X_1 \cdot 2 \cdot 25 \cdot 8 \cdot 0,02 + X_2 \cdot 2 \cdot 15 \cdot 8 \cdot 0,05 = 8 X_1 + 12 X_2;$$

$$Y = \text{ЗП} + \text{ЗБ} = 32 X_1 + 24 X_2 + 8 X_1 + 12 X_2 = 40 X_1 + 36 X_2.$$

Лінії рівня ЦФ зображені пунктиром на рис. 4.18.

Обмеження:

- на кількість контролерів $X_1 \leq 8$ (лінія BA на рис. 4.18); $X_2 \leq 10$ (лінія CB на рис. 4.18);
- на змінну норму ВТК $X_1 \cdot 8 \cdot 25 + X_2 \cdot 8 \cdot 15 = 200 X_1 + 120 X_2 \geq 1800$ або $5 X_1 + 3 X_2 \geq 45$ (лінія CA на рис. 4.18).

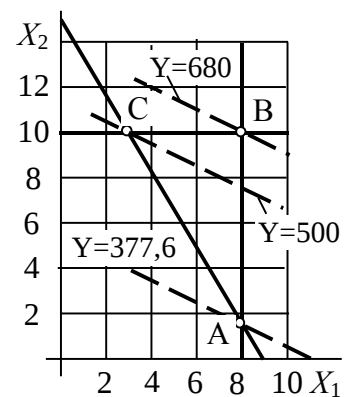


Рис. 4.18

Як видно, оптимальною точкою є т. А з координатами $X_1=8$, $X_2=1\frac{2}{3}$. Це означає, що один з контролерів III розряду має працювати дві третини робочого дня.

З цього прикладу можна бачити загальні властивості задач ЛП:

- лінійна ЦФ отримує умовний екстремум лише внаслідок наявності обмежень;
- умовний екстремум завжди розташований на межі допустимої області (на грані або найчастіше у вершині багатокутника).

Чисельні методи задач ЛП звичайно базуються на послідовному переборі вершин багатогранника, створеного гіперплощинами обмежень в n -вимірному факторному просторі. Внаслідок лінійності ЦФ та обмежень швидкодія цих методів достатньо велика.

До типових задач, що розв'язуються методами ЛП, належать:

- оптимізація обсягів виробництва за обмежених запасів сировини;
- оптимізація завантаження наявного технологічного обладнання;
- оптимізація складу суміші із заданими властивостями;
- оптимізація кадрового складу колективу або графіка його роботи;
- оптимальний вибір списку постачальників;
- транспортні задачі (розподілення потоків у транспортній мережі або вибір оптимального маршруту транспортування).

Критерієм оптимізації для більшості задач ЛП є мінімум витрат або максимум прибутку.

Розглянемо два варіанти оптимізації передачі матеріальних, енергетичних або інформаційних ресурсів транспортною мережею заданої конфігурації. Є мережа (рис. 4.19), яка складається з вузлів ($A, \dots, F$) та гілок, що їх з'єднують. Для кожної гілки відомі:

- пропускна спроможність Q_{mi} ;
- вартість переміщення одиниці потоку C_i .

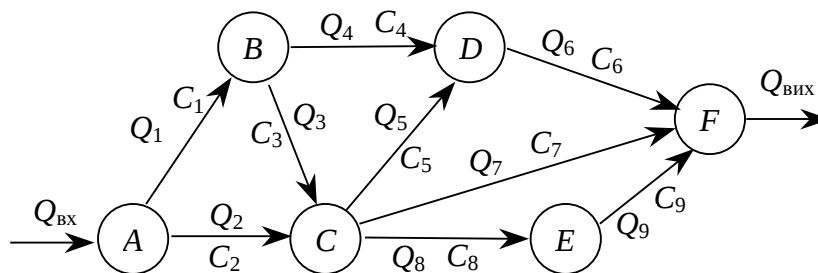


Рис. 4.19 Транспортна мережа

Умова безперервності потоку у вузлі $\sum_{j=1}^k Q_j = 0$ (j – номер гілки, зв'язаної з

даним вузлом, k – їхня кількість) виступає як обмеження. Ця умова означає, що в жодному вузлі не відбувається накопичення транспортованих ресурсів. Мережа може мати кілька входів та виходів, але сума вхідних потоків завжди дорівнює сумі вихідних: $\sum Q_{\text{ВХ}} = \sum Q_{\text{ВІХ}}$.

Варіант 1. Вважаючи потоки на вході та виході заданими ($Q_{\text{вх}} = Q_{\text{вих}} = Q$), знайти розподіл потоків в мережі, який мінімізує витрати на транспортування.

Цільова функція визначається як сума вартостей транспортування всіма гілками: $Y = \sum_{i=1}^n C_i Q_i$ (тут i – номер гілки). Як фактори виступають потоки Q_i . Обмеження з боку умови безперервності потоку у кожному з вузлів:

- вузол А: $Q - Q_1 - Q_2 = 0$;
- вузол В: $Q_1 - Q_3 - Q_4 = 0$;
- вузол С: $Q_2 + Q_3 - Q_5 - Q_6 - Q_7 = 0$;
- вузол D: $Q_4 + Q_5 - Q_6 = 0$;
- вузол Е: $Q_8 - Q_9 = 0$;
- вузол F: $Q_6 + Q_7 + Q_9 - Q = 0$.

Обмеження за пропускною спроможністю та напрямом передачі (останнє не завжди має місце):

$$0 \leq Q_i \leq Q_{mi}.$$

Варіант 2. Визначити розподіл потоків, який забезпечує максимум вихідного потоку. Обмеження ті ж самі, що й у попередньому варіанті, але з одного з них (наприклад, першого) можна виразити ЦФ: $Q = Q_1 + Q_2 \Rightarrow \max$.

Застосування транспортних задач ЛП:

- перевезення вантажів (пасажирів) мережею автомобільних доріг або залізницею;
- передача інформації каналами зв'язку (провідний зв'язок, Інтернет тощо);
- передача активної та реактивної енергії енергосистемою;
- розподілення струмів у складному електричному колі (у тому числі нелінійному);
- вентиляція розгалужених підземних виробок тощо.

5 ЕКСТРЕМАЛЬНІ СИСТЕМИ РЕГУЛЮВАННЯ

5.1 Загальна характеристика екстремальних систем регулювання

Екстремальні системи регулювання (ЕСР) належать до класу статичних адаптивних самонастроювальних. Вони оптимізують певну статичну характеристику об'єкту керування.

Як і всі самонастроювальні системи, ЕСР мають спеціальний контур самонастроювання, за допомогою якого оцінюється та змінюється певний вихідний параметр Y (цільова функція), який має екстремальну залежність від керуючих впливів X_i . Прикладами цільових функцій ЕСР можуть слугувати:

- коефіцієнт корисної дії установки як функція ступеню навантаження;
- струм статора, потужність втрат та $\cos\phi$ асинхронного двигуна як функція рівня живильної напруги за незмінних частоти живлення та моменту навантаження;
- кількість теплоти, що виділяється у топці, як функція витрат палива та повітря;
- потужність, генерована вітрогенератором, як функція кута повороту лопаті вітротурбіни;
- вихідна потужність сонячної батареї, як функція вихідного струму тощо.

ЕСР використовують тоді, коли положення екстремуму ЦФ (тобто, оптимальні значення факторів X_i) під впливом збурень змінюється у часі (інакше достатньо було б визначити це положення одноразово та використати надалі систему автоматичної стабілізації).

У процесі роботи ЕСР виконують наступні задачі:

- організації пошуку екстремуму ЦФ;
- локалізації цього екстремуму із заданою точністю;
- утримання знайденого екстремуму в умовах дії збурень.

Вимоги до ЕСР:

- стійкість (процес пошуку має з часом зійтися до певного інтервалу локалізації екстремуму);
- швидкодія (тривалість пошуку повинна бути достатньо малою);
- точність (знайдений інтервал локалізації має бути достатньо вузьким).

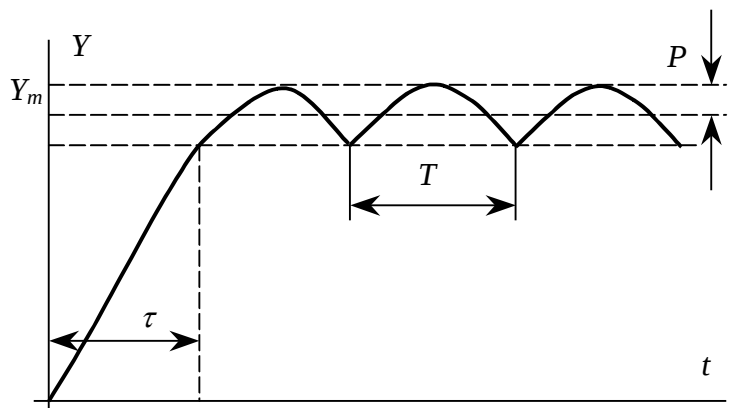


Рис. 5.1 Зміна у часі цільової функції ЕСР

На рис. 5.1 показана зміна у часі цільової функції ЕСР. Тривалість виходу до околиць екстремуму τ характеризує швидкодію ЕСР, діапазон коливань ЦФ P після локалізації екстремуму (втрати на пошук) — точність. Процес утримання екстремуму завжди має автоколивальний характер з періодом T .

Методи пошуку екстремуму в ЕСР практично такі ж самі, як і нелінійному програмуванні, проте термінологія не завжди співпадає. Переважно застосовують два методи пошуку:

- пошук за прирощенням Y (аналогія прямих методів, розглянутих у розділі 4: напрям руху обирають у результаті порівняння останніх значень ЦФ);
- пошук за похідною Y (аналогія градієнтних методів пошуку).

5.2 Структури екстремальних систем керування

5.2.1 Система з пошуком за прирощенням

Приклад одновимірної ЕСР із запам'ятовуванням екстремуму зображено на рис. 5.2. До входу об'єкту керування ОК від виконавчого елементу ВЕ подається сигнал $X = \int U dt$, який є інтегралом від сигналу керування U . Останній вано релейним елементом із гістерезисом РЕ (регулятором), вихідний сигнал якого може мати лише два незмінних рівня: позитивний та негативний. Елемент пам'яті ЕП запам'ятовує поточне значення вихідної змінної об'єкта керування Y

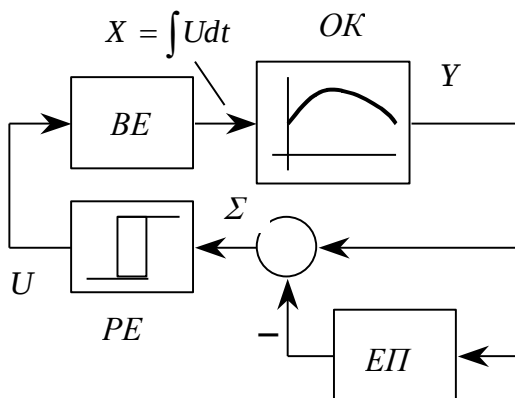


Рис. 5.2 ЕСР із запам'ятовуванням екстремуму

та подає її до входу суматора Σ на ступному такті Δt роботи системи. Тому вихідний сигнал суматора, який надходить до входу регулятора, порційний поточному прирощенню ЦФ. Якщо ЦФ на поточному кроці льша, ніж на попередньому, вихідний сигнал регулятора є позитивним, а сигнал X на вході ОК лінійно зростає. Після досягнення екстремуму ЦФ її наступне значення стане меншим за попереднє, знак сигналу регулятора зміниться, сигнал на вході ОК почне зменшуватися, доки система знову не знайде екстремум. На рис. 5.3 зображено процес пошуку та утримання екстремуму в системі з керуванням за прирощенням. Кожного разу, коли значення ЦФ Y стає меншим від попереднього, знак сигналу регулятора стає від'ємним, сигнал виконавчого елементу зменшується. Після локалізації муму система працює переважно за точкою екстремуму A_0 .

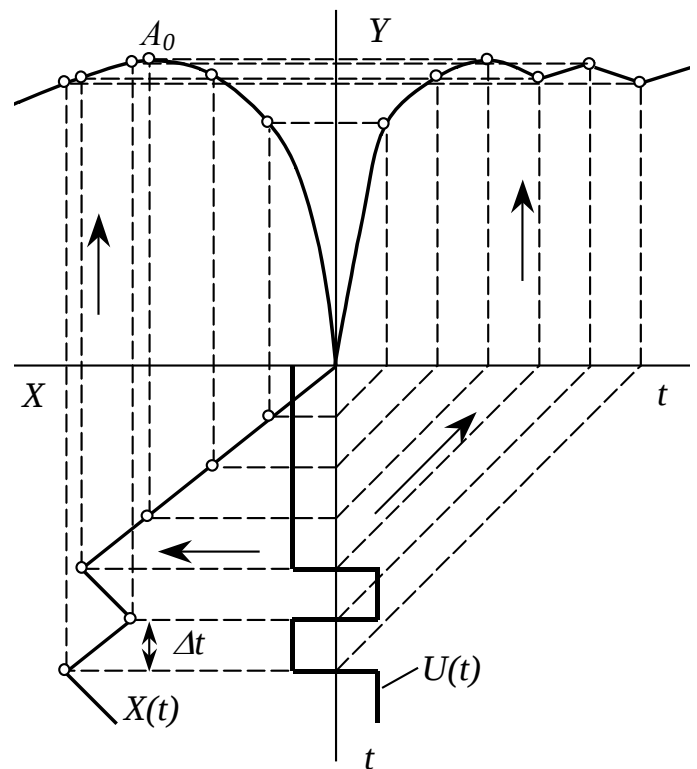


Рис. 5.3 Процес пошуку в ЕСР за рис. 5.2

5.2.2 Системи з пошуком за похідною

Екстремальну систему з пропорційним керуванням за похідною зображено на рис. 5.4. Два блоки диференціювання Д1 та Д2 визначають поточні значення похідних за часом вхідного $\frac{dX}{dt}$ та вихідного $\frac{dY}{dt}$ сигналів об'єкту керування. На виході блоку ділення маємо похідну $\frac{dY}{dX} = \frac{dY/dt}{dX/dt}$, яка після під-

силення підсилювачем П надходить до виконавчого елементу. Завдяки цьому темп руху до екстремуму пропорційний поточній похідній $\frac{dY}{dX}$ з урахуванням її знаку. Після досягнення екстремуму знак похідної змінюється, що призводить до зміни напрямку руху.

Система, наведена на рис. 5.5, дещо простіша завдяки тому, що використано лише один блок диференціювання Д, а регулятор замість пропорційного – релейний (РЕ).

Більшого поширення набули ЕСР з син-

хронним детектуванням, в яких темп руху також пропорційний похідній $\frac{dY}{dX}$ (рис. 5.6). У них відсутні блоки диференціювання, а похідну визначають опосередковано, за допомогою модуляції/демодуляції. Принцип дії системи ґрунтується на додаванні до корисного вхідного сигналу X синусоїдного пробного сигналу X_n від генератора ГПС, який трохи “розхи-

тує” систему. Амплітуду та частоту пробного сигналу обирають такими, щоб у вихідній змінній Y з'явилися коливання відповідної частоти, проте не настільки великі, щоб дезорганізувати роботу ОК. Детектор фази ДФ (фактично блок множення) разом із фільтром низьких частот Φ формують сигнал, пропорційний $\frac{dY}{dX}$. Математичне підґрунтя роботи такої системи видно з наступного.

Нехай статична характеристика ОК визначається квадратичною залежністю $Y = Y_0 + a(X - X_0)^2$,

звідки похідна

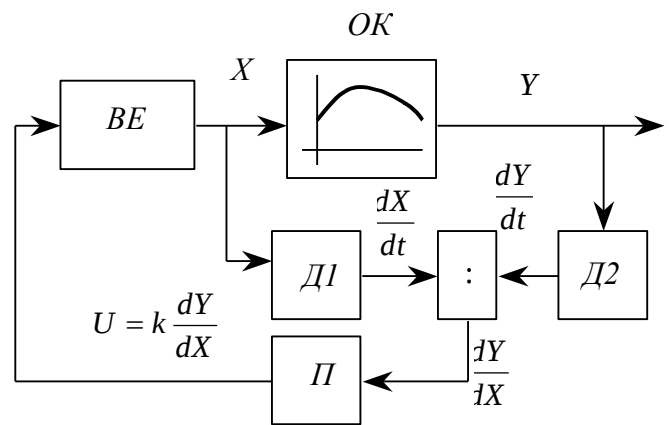


Рис. 5.4 Пропорційна ЕСР із пошуком за похідною

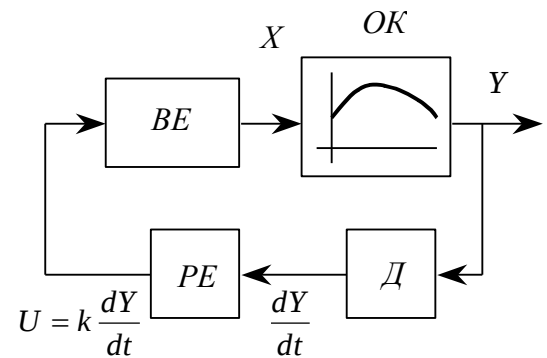


Рис. 5.5 Релейна ЕСР із пошуком за похідною

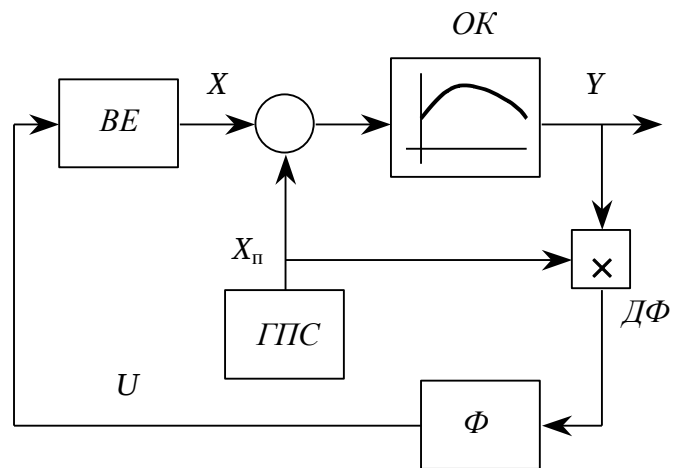


Рис. 5.6 ЕСР із синхронним детектуванням

$$\frac{dY}{dX} = 2a(X - X_0).$$

Якщо додати до корисного вхідного сигналу пробний $X_n = A \sin \omega t$, отримаємо трохи інші значення ЦФ:

$$\begin{aligned} Y' &= Y(X + X_n) = Y_0 + a(X - X_0 + A \sin \omega t)^2 = \\ &= Y_0 + a[(X - X_0)^2 + A^2 \sin^2 \omega t + 2(X - X_0)A \sin \omega t] = Y(X) + aA^2 \sin^2 \omega t + 2a(X - X_0)A \sin \omega t = \\ &= Y(X) + aA \sin^2 \omega t + \frac{dY}{dX} A \sin \omega t. \end{aligned}$$

Якщо помножити останній вираз на пробний сигнал, матимемо

$$Y' \cdot X_n = Y(X)A \sin \omega t + aA^2 \sin^3 \omega t + \frac{dY}{dX} A^2 \sin^2 \omega t.$$

Фільтр низьких частот фактично виділяє середнє (за період пробного сигналу) значення вихідного сигналу детектора фази. Середні за період значення першої та другої складових останнього виразу дорівнюють нулю, оскільки пропорційні відповідно першій та третій ступеням синусу. Оскільки ж $\frac{1}{2\pi} \int_0^{2\pi} \sin^2 \omega t d(\omega t) = 0,5$, то вихідний сигнал фільтру дорівнюватиме

$$U = \frac{1}{2\pi} \int_0^{2\pi} Y' \cdot X_n d(\omega t) = \frac{dY}{dX} \cdot \frac{A^2}{2}.$$

Таким чином, темп руху до екстремуму буде пропорційним похідній $\frac{dY}{dX}$, так само, як і в системі за рис. 5.4.

5.3 Особливості багатовимірних систем екстремального керування

Якщо ЦФ об'єкту керування одночасно залежить від кількох керуючих впливів, для пошуку її екстремуму необхідно мати багатовимірну екстремальну систему регулювання. У цих системах можливі дві стратегії: пошук з одночасною зміною всіх керуючих впливів, або з почерговою.

На рис. 5.7 показано двовимірну екстремальну систему першого типу з використанням принципу синхронного детектування. Вона має два однакові канали керування для формування сигналів X_1, X_2 . Частоти пробних сигналів мають бути неоднаковими. У разі цифрової реалізації такої системи може бути використаний будь-який з багатовимірних методів пошуку, розглянутих у розділі 4.

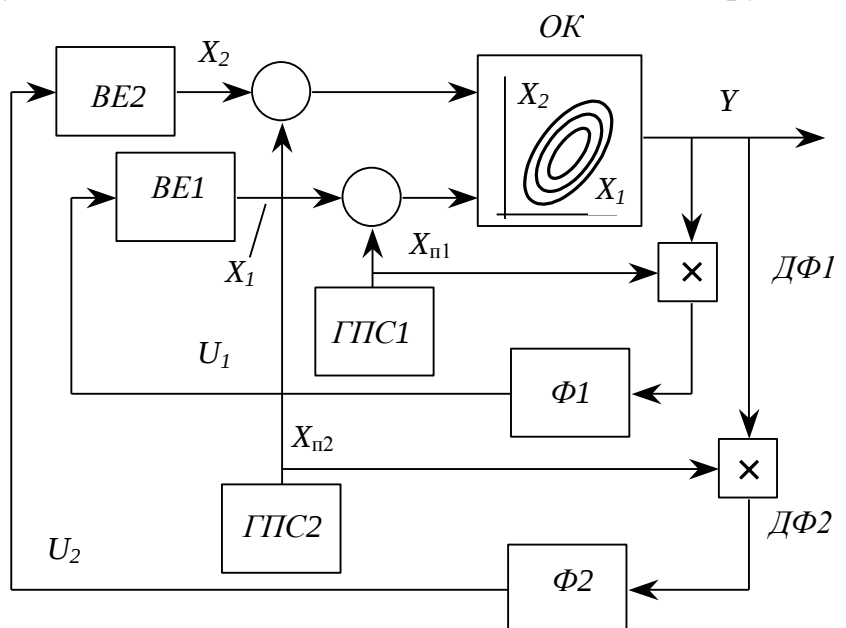


Рис. 5.7 Двовимірна ЕСР
з одночасною зміною керуючих впливів

Якщо ЕСР має лише один канал керування (як на рис. 5.8), у ній можливий тільки пошук за методом Гауса-Зайделя. Логічний пристрій ЛП після зміни знаку сигналу на виході фільтру Φ (тобто після чергового переходу через екстремум) перемикає канали керування, забезпечуючи почергові рухи вздовж осей координат. Так само можна організувати ЕСР із перемиканням каналів, використовуючи структури рис. 5.2, 5.4, 5.5.

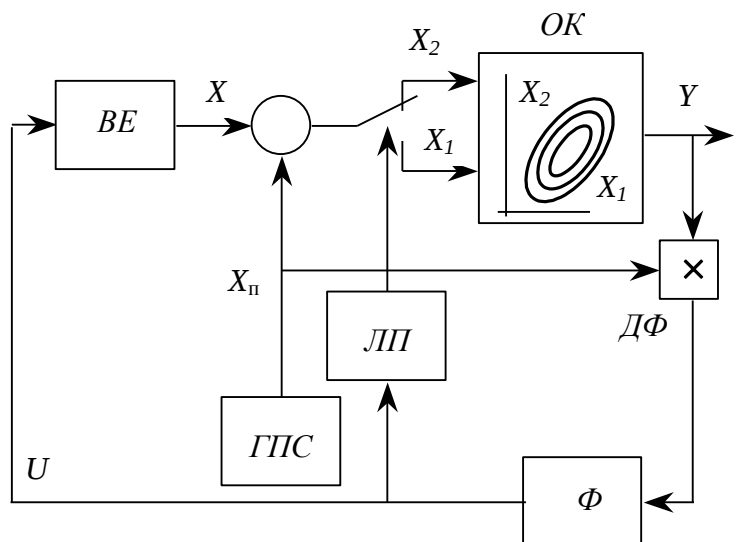


Рис. 5.8 Двовимірна ЕСР
із перемиканням каналів керування

Література

1. Уайльд Д.Дж. Методы поиска экстремума. – М.: Наука, 1967. – 267 с.
2. Химмельблау Д. Прикладное нелинейное программирование. – М.: Мир, 1975. – 536 с.
3. Реклейтис Г., Рейвиндран А., Рэгсдел К. Оптимизация в технике: В 2-х кн. – М.: Мир, 1986. Кн.1. – 350 с.; Кн.2. – 320 с.
4. Ивоботенко Б.А., Ильинский Н.Ф., Копылов И.П. Планирование эксперимента в электромеханике. – М.: Энергия, 1975. – 184 с.
5. Адлер Ю.П., Маркова Е.В., Грановский Ю.В. Планирование эксперимента при поиске оптимальных условий. – М.: Наука, 1971. – 283 с.
6. Уайльд Д. Оптимальное проектирование. – М.: Мир, 1981. – 272 с.
7. Первозванский А.А. Поиск. – М.: Наука, 1970. – 264 с.
8. Моисеев Н.Н., Иванилов Ю.П., Столярова Е.М. Методы оптимизации. – М.: Наука, 1978. – 352 с.
9. Банди Б. Методы оптимизации: Вводный курс. – М.: Радио и связь, 1988. – 128 с.
10. Электропривод летательных аппаратов/ В.А.Полковников, Б.И.Петров, Б.Н.Попов и др. – М.: Машиностроение, 1990. – 352 с.
11. Олейников В.А., Зотов Н.С., Пришвин А.М. Основы оптимального и экстремального управления. – М.: Высшая шк., 1969. – 296 с.