4.2 Численные методы безусловной минимизации функции многих переменных

Ставится задача минимизации функции $f(\boldsymbol{x}) = f(x_1, x_2, ...x_n)$ в некоторой замкнутой области. Пусть $\boldsymbol{a} = (a_1, a_2, ..., a_n)$ точка минимума $f(\boldsymbol{x})$. Будем говорить, что с точностью до ε точка $\boldsymbol{x}$ может быть взята в качестве приближенного значения точки минимума, если

$\| \boldsymbol{a} - \boldsymbol{x} \| < \varepsilon$, где $\| \boldsymbol{a} - \boldsymbol{x} \| = \sqrt{\sum_{i=1}^{n} (a_i - x_i)^2}$ или

$\| \boldsymbol{a} - \boldsymbol{x} \| = max \left| a_i - x_i \right| \quad (i = \overline{1,n})$.

Для геометрической иллюстрации методов будем использовать функцию двух переменных. Напомним, что линиями уровня функции $Z=f(x,y)$ называют множество точек (x, y), удовлетворяющих уравнению $f(x,y)=c$. Меняя c, мы получаем различные линии уровня функции $f(x,y)$. Геометрически линия уровня – это проекция на плоскость XOY линии пересечения $Z=f(x,y)$ и плоскости $Z=c$. Имея множество линий уровня, мы получаем представление о поведении $Z=f(x,y)$, говорят о рельефе функции $Z=f(x,y)$.

4.2.1 Методы многомерного прямого поиска

Суть методов многомерного прямого поиска, изложенных ниже, в том, что выбирают некоторую точку $\boldsymbol{x}^{(0)} = (x_1^{(0)}, ..., x_n^{(0)})$ и допустимое направление поиска $\boldsymbol{d}$. Затем, отправляясь от точки $\boldsymbol{x}^{(0)}$ в направлении $\boldsymbol{d}$, минимизируют функцию одного переменного λ: $f(\boldsymbol{x} + \lambda \boldsymbol{d})$, изложенными выше методами.

Найдя $\lambda^{(0)}$, при котором $f(\boldsymbol{x}^{(0)} + \lambda \boldsymbol{d})$ получает минимальное значение, мы тем самым нашли точку $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \lambda^{(0)} \boldsymbol{d}$, в которой значение f, вообще говоря, меньше чем в точке $\boldsymbol{x}^{(0)}$. Далее, отправляясь от $\boldsymbol{x}^{(1)}$ в некотором новом направлении $\boldsymbol{d}_1$, получаем некоторую точку $\boldsymbol{x}^{(2)}$, в которой значение f вообще говоря, меньше чем

в $\boldsymbol{x}^{(1)}$ и т.д. Возникают вопросы: 1) Как целесообразнее выбирать $\boldsymbol{d}_i$? 2) Сходится ли последовательность $\boldsymbol{x}^{(0)}$, $\boldsymbol{x}^{(1)}$, . . .? 3) Как оценить погрешность?

4.2.2 Метод циклического покоординатного спуска

Опишем алгоритм одного из таких методов, выбирая в качестве направлений поиска координатные векторы $\boldsymbol{d}_1 = \boldsymbol{l}_1, \boldsymbol{d}_2 = \boldsymbol{l}_2, ..., \boldsymbol{d}_n = \boldsymbol{l}_n$, т.е. $\boldsymbol{d}_i = \boldsymbol{l}_i$ - вектор, все компоненты которого, за исключением i равны нулю.

Тогда

$$\boldsymbol{x}^{(0)} + \lambda \boldsymbol{l}_i = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_{i-1}^{(0)}, x_i^{(0)} + \lambda, x_{i+1}^{(0)}, \ldots, x_n^{(0)}\right),$$

$$f(\boldsymbol{x}^{(0)} + \lambda \boldsymbol{l}_i) = f\left(x_1^{(0)}, x_2^{(0)}, ..., x_{i-1}^{(0)}, x_i^{(0)} + \lambda, x_{i+1}^{(0)}, ..., x_n^{(0)}\right) \quad \text{и,}$$

следовательно, при минимизации функции $f(\boldsymbol{x}^{(0)} + \lambda \boldsymbol{l}_i)$, речь идет о минимизации функции одной переменной $x_i = x_i^{(0)} + \lambda_i$, при фиксированных остальных переменных.

Выберем начальную точку $\mathbf{y}_1 = \boldsymbol{x}^{(0)}$, направление $\boldsymbol{d}_1 = \boldsymbol{l}_1$ и, минимизируя $f(\mathbf{y}_1 + \lambda \boldsymbol{l}_1)$, найдем, что минимум этой функции достигается при λ_1, в точке $\mathbf{y}_2 = \mathbf{y}_1 + \lambda_1 \boldsymbol{l}_1 = \left(x_1^{(0)} + \lambda_1, x_2^{(0)}, ..., x_n^{(0)}\right)$. Положим $\mathbf{y}_2 = \mathbf{y}_1 + \lambda_1 \boldsymbol{l}_1$, выберем направление $\boldsymbol{l}_2$ и, минимизируя $f(\mathbf{y}_2 + \lambda \boldsymbol{l}_2)$ найдем, что минимум этой функции достигается при λ_2 в точке $\mathbf{y}_3 = \mathbf{y}_2 + \lambda_2 \boldsymbol{l}_2 = \left(x_1^{(0)} + \lambda_1, x_2^{(0)} + \lambda_2, x_3^{(0)}, ..., x_n^{(0)}\right)$.

После n шагов найдем

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \lambda_n \boldsymbol{l}_n = \left(x_1^{(0)} + \lambda_1, x_2^{(0)} + \lambda_2, ..., x_n^{(0)} + \lambda_n\right).$$

Положим $\boldsymbol{x}^{(1)} = \mathbf{y}_{n+1}$ и тем самым завершим один цикл покоординатного спуска.

Отправляясь от $\boldsymbol{x}^{(1)}$, как от начальной точки, можем найти $\boldsymbol{x}^{(2)}$ и т.д.

Процесс завершить, если

$$\| \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} \| < \varepsilon,$$

где ε – требуемая точность,

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| = \sqrt{\sum_{i=1}^{n} \left(x_i^{(k+1)} - x_i^{(k)} \right)^2} \text{ или}$$

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| = \max| x_i^{(k+1)} - x_i^{(k)} | \, (i = \overline{1,n}).$$

Проиллюстрируем метод с помощью функции двух переменных (рисунок 4.4).

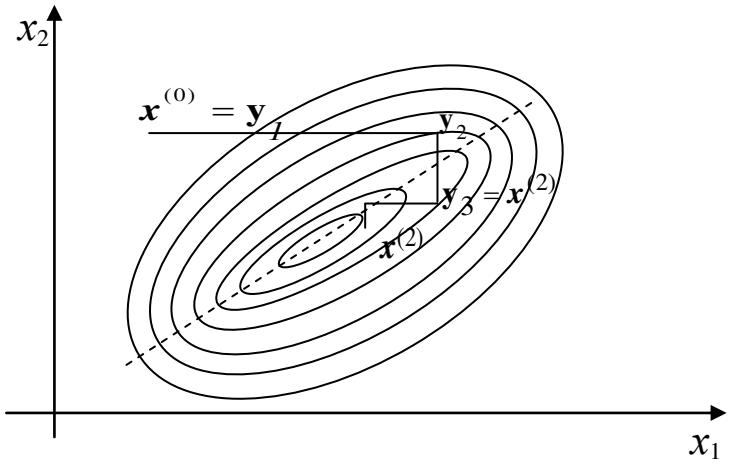

Рисунок 4.4 – Метод циклического покоординатного спуска

Двигаясь от точки $\boldsymbol{x}^{(0)} = \mathbf{y}_1$ параллельно оси OX_1, мы переходим от линий уровня с большим значением $f(x_1, x_2) = c$ к меньшим. Точка $\mathbf{y}_2$, в которой мы коснемся некоторой линии уровня, является точкой минимума функции $f(x_1, x_2^{(0)})$ в направлении, параллельном оси OX_1. Двигаясь от точки $\mathbf{y}_2$ параллельно оси OX_2, мы, переходя от линий уровня с большим "c" к линиям уровня с меньшим "c", достигнем минимального значения в точке $\mathbf{y}_3$ - точке касания некоторой линии уровня. В точке $\mathbf{y}_3$ завершается цикл покоординатного спуска и получаем $\boldsymbol{x}^{(1)} = \mathbf{y}_3$.

Отправляясь от $\boldsymbol{x}^{(1)}$, как от начальной точки, можем найти $\boldsymbol{x}^{(2)}$ и т.д. Остается указать условия сходимости последовательности $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}$, ... и указать оценку погрешности.

Для сходимости метода циклического покоординатного спуска достаточно следующих требований:

1) минимум $f(\boldsymbol{x})$ вдоль любого направления единственен;

2) последовательность точек $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(k)}$ принадлежит некоторому замкнутому ограниченному подмножеству области $\boldsymbol{D}$.

Что касается погрешности, то ее можно определить по формуле: $\| \boldsymbol{a} - \boldsymbol{x}^{(k+1)} \| \leq \| \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} \|$.

ЗАМЕЧАНИЕ. Если функция f не является дифференцируемой в некоторых точках, то метод может остановиться в неоптимальной точке (рисунок 4.5).

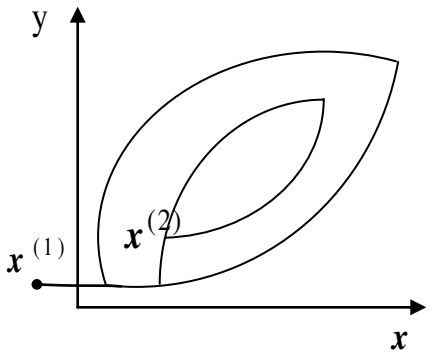

Рисунок 4.5 – Неоптимальная точка

Ясно, что поиск вдоль любой координатной оси в точке $\boldsymbol{x}^{(2)}$ не приводит к улучшению целевой функции. Это вызвано наличием так называемого оврага, то есть точки недифференцируемой функции $f(\boldsymbol{x})$. Разрешить эту ситуацию можно используя, например, метод Хука−Дживса.

4.2.3 Метод Хука-Дживса

Метод Хука-Дживса осуществляет два типа поиска: исследующий поиск и поиск по образцу. Выбрав начальную точку $\boldsymbol{x}_1$ методом циклического покоординатного спуска, находим точку $\boldsymbol{x}_2$, т.е. осуществляем исследующий поиск. Если $\| \boldsymbol{x}_2 - \boldsymbol{x}_1 \| < \varepsilon$, то точка минимума найдена, иначе осуществляем поиск по образцу в направлении $\boldsymbol{x}_2 - \boldsymbol{x}_1 = \boldsymbol{d}$,

что приводит нас в некоторую точку **y**. Приняв **y** за $\boldsymbol{x}_1$ вновь проводим исследующий поиск и т.д. Сходимость метода Хука-Дживса обеспечена при тех же условиях, что и покоординатного спуска. Иллюстрация метода на рисунке 4.6 .

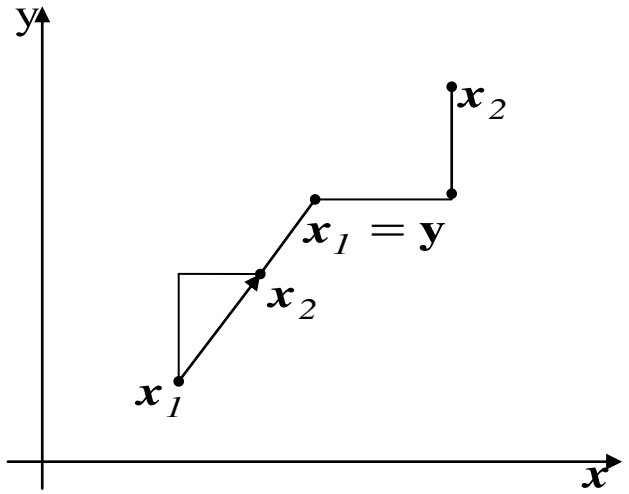

Рисунок 4.6 – Метод Хука-Дживса

4.2.4 Метод наискорейшего спуска

Известно, что направлением наибольшего возрастания функции f в точке $\boldsymbol{x}^{(0)}$ является направление, задаваемое вектором $grad\ f(\boldsymbol{x}^{(0)})$, а $(-grad\ f(\boldsymbol{x}^{(0)}))$ задает направление наибольшего убывания функции f в точке $\boldsymbol{x}^{(0)}$. Учитывая это, следует осуществлять линейный поиск не в направлении осей координат, а в направлении $-grad\ f$.

Рассмотрим алгоритм метода.

1. Выбираем $\boldsymbol{x}_i$, $i=1$.

2. Если $\| grad\ f(\boldsymbol{x}_i)\| < \varepsilon$, то $\boldsymbol{x}_i$- искомая точка.

3. В противном случае положим $d_i = -grad\ f(\boldsymbol{x}_i)$ и решив задачу о минимуме $f(\boldsymbol{x}_i + \lambda d_i)$, найдем $\lambda_i > 0$.

4. Найдем $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \lambda_i d_i$ и переходим к пункту 2, положив $i = i+1$.

Сходимость метода обеспечена, если f непрерывно дифференцируема, а генерируемая последовательность, принадлежит замкнутому ограниченному множеству.

Недостатком метода наискорейшего спуска является зачастую медленная сходимость в окрестности стационарной точки. Это очевидно, например, в тех случаях, когда линии уровня вытянуты в окрестности оптимальной точки. О функциях, поверхности уровня которых сильно вытянуты, говорят, что она имеет "овражный" характер. Геометрически медленная сходимость объясняется так: спустившись на "дно оврага" мы, двигаясь в направлении ($-grad\ f$), будем переходить с одного склона оврага на другой, то есть зигзагообразно продвигаться к точке минимума (рисунок 4.7).

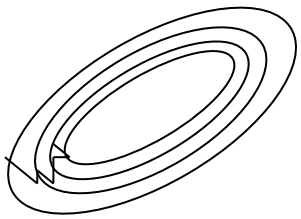

Рисунок 4.7 – Овражная функция

Что касается степени "овражности", то ее можно охарактеризовать с помощью минимального (λ_{min}) и максимального (λ_{max}) собственных чисел матрицы Гессе – $\left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)(i=\overline{1,n};\ \ j=\overline{1,n})$. Чем меньше λ_{min} / λ_{max}, тем больше овражность. Более предпочтительными в таких случаях являются методы Ньютона и сопряженных направлений.

Коротко суть первого из них состоит в том, что функция $f(\boldsymbol{x})$ аппроксимируется (с помощью формулы Тейлора) многочленами второй степени, для которых находятся точки минимума. Последовательность таких точек приводит при определенных условиях к искомой точке минимума $f(\boldsymbol{x})$. Неудобством метода является необходимость многократного обращения матрицы Гессе.

Второй из методов для получения очередной точки требует проведения последовательной минимизации по каждому из n, специальным образом построенных направлений, которые называют сопряженными направлениями. Для квадратичной функции минимизация

вдоль "n" таких направлений позволяет "точно" достичь точки минимума, а следовательно можно ожидать хороших результатов и для достаточно гладких функций. Подробнее с этими методами можно познакомиться в [4].