К. В. Григорьева

Методические указания

Тема 3. Методы решения задачи минимизации квадратичной функции

Факультет ПМ-ПУ СПбГУ 2007 г.

ОГЛАВЛЕНИЕ

1. ПОСТАНОВКА ЗАДАЧИ. ВСПОМОГАТЕЛЬНЫЕ СВЕДЕНИЯ 2
2. СВОЙСТВА КВАДРАТИЧНОЙ ФУНКЦИИ 3
3. МЕТОДЫ СПУСКА → ГРАДИЕНТНЫЕ МЕТОДЫ → МЕТОД НАИСКОРЕЙШЕГО ГРАДИЕНТНОГО СПУСКА 4
3.1. Общий план методов спуска 4
3.2. Выбор направления и шага спуска. Метод наискорейшего спуска 4
3.3. Метод наискорейшего градиентного спуска 5
3.4. Критерий окончания итераций 5
3.5. Геометрический смысл МНГС 5
3.6. Сходимость и проблема "оврагов" 5
3.7. Примеры 6
4. МЕТОД ПОКООРДИНАТНОГО СПУСКА 7
4.1. Геометрическая интерпретация 7
4.2. Примеры 7
5. МЕТОД СОПРЯЖЕННЫХ ГРАДИЕНТОВ 8
5.1. Алгоритм метода 8
5.2. Геометрическая интерпретация 9
5.3. Примеры 9
6. ТИПОВОЙ РАСЧЕТ 9
ЛИТЕРАТУРА 9

1. ПОСТАНОВКА ЗАДАЧИ. ВСПОМОГАТЕЛЬНЫЕ СВЕДЕНИЯ

Пусть R^n – n-мерное евклидово пространство.

Определение 1.1. ***Задачей нелинейного программирования*** называется задача минимизации функции $f:\ (D \subset R^n) \to R^1$ при условии $x \in X \subset R^n$. Ее стандартное обозначение

$$f(x) \to \min_{x \in X \cap D}. \tag{1.1}$$

Функция f называется целевой или минимизируемой функцией. Если $X = R^n$, то тогда (1.1) называется ***задачей безусловной минимизации***. Если $X \neq R^n$, то – ***задачей условной минимизации***.

Определение 1.2. ***Решением задачи*** **(1.1)** будем называть либо пару (x^*, f_*), если существует конечная точка минимума x^*, либо пару $\left(\{x^k\}_{k=1}^{\infty}, f_*\right)$, где $\{x^k\}_{k=1}^{\infty} \subset X$ – минимизирующая последовательность со свойством $f(x^k) \xrightarrow[k\to\infty]{} f_* := \inf\{f(x), x \in X\}$.

Конечная точка минимума существует не всегда. Например, для выпуклой функции $f(x) = e^x$ существует конечный инфимум $f_* = 0$ и минимизирующая последовательность $x_i = -i$, но нет конечной точки минимума. Бывает, что нет и конечного инфимума. Например, если $f(x) = \ln(x),\ D = (0, +\infty)$, то инфимум $f_* = -\infty$ доставляется минимизирующей последовательностью $x_i = 1/i$. Минимизирующая последовательность существует всегда.

Определение 1.3. Точка $x^* \in X$ называется ***точкой глобального минимума*** функции f на множестве X, если для всех $x \in X$ выполняется неравенство

$$f(x^*) \le f(x). \tag{1.2}$$

Значение $f(x^*)$ называется минимальным значением функции f на X. Точка $x^* \in X$ называется ***точкой локального минимума*** функции f, если существует такая δ-окрестность X_δ этой точки, что для всех $x \in X_\delta$ выполняется неравенство (1.2). Если имеет место $f(x^*) < f(x)$ для всех $x \in X \setminus \{x^*\}$ или для всех $x \in X_\delta \setminus \{x^*\}$, то x^* – точка ***строгого глобального или локального минимума***, соответственно.

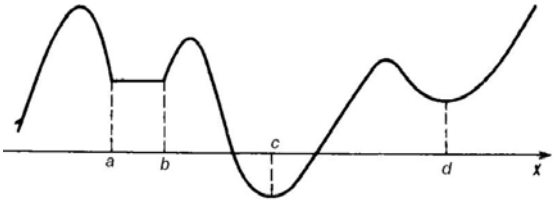

Рис. 1.1. Глобальный (c), строгий локальный (d) и нестрогие локальные $x \in [a, b]$ минимумы.

В нелинейном программировании особую роль играют выпуклые функции, так как локальный минимум выпуклой функции является глобальным минимумом, а строгая выпуклость обеспечивает еще единственность глобального минимума.

Определение 1.4. Функция f называется ***выпуклой*** на множестве X, если $\forall\ x, \bar{x} \in X,\ \lambda \in (0,1)$ выполняется неравенство

$$f(\lambda x + (1-\lambda)\bar{x}) \le \lambda f(x) + (1-\lambda) f(\bar{x}). \tag{1.3}$$

Если в (1.3) заменить $\le$ на $<$, то получим определение строго выпуклой функции.

Геометрический смысл этого определения — график функции f на отрезке выпуклости $[x, \bar{x}]$ лежит ниже хорды, соединяющей точки $(x, f(x))$ и $(\bar{x}, f(\bar{x}))$ (см. рис. 1.2).

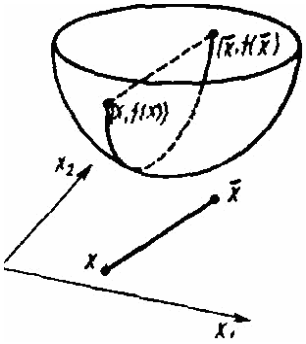

Рис. 1.2.

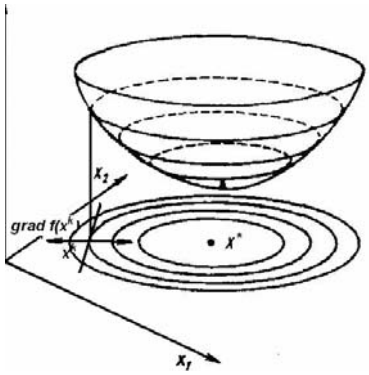

Рис. 1.3.

Здесь будет рассмотрена задача безусловной минимизации квадратичной функции в пространстве

столбцов R^n. Напомним следующие определения.

Определение 1.5. Вектор-столбец из первых частных производных функции $f(x)$

$$f'(x) := \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)^T$$

называется ***градиентом***, а вектор-столбец $g(x) = -f'(x)$ называется ***антиградиентом***.

Определение 1.6. Точка x, для которой выполняется равенство $f'(x) = 0$, называется ***стационарной точкой*** функции f.

Определение 1.7. Множество точек Γ_c, для которых целевая функция принимает постоянное значение $f(x) = c$, в случае $n > 2$ называется ***поверхностью уровня***, а в случае $n = 2$ – ***линией уровня***.

Например, функция $z = f(x_1, x_2)$ задает в трехмерном пространстве некоторую поверхность, низшая точка которой, согласно определению 1.2, есть решение задачи минимизации (см. рис. 1.3). Проведем несколько плоскостей вида $z = const$. Линиями уровня будут проекции на плоскость Ox_1x_2 линий пересечения этих плоскостей с поверхностью.

Теорема 1.1. Если в точке x градиент не равен нулю, то он перпендикулярен к проходящей через эту точку поверхности уровня.

В направлении антиградиента происходит наискорейшее убывание функции, т. е. если мы делаем шаг бесконечно малой длины в некотором направлении (см. рис. 1.3), то наибольшее убывание функции произойдет тогда, когда этот шаг совпадет по направлению с антиградиентом. Это свойство антиградиента лежит в основе ряда итерационных методов минимизации функций.

2. СВОЙСТВА КВАДРАТИЧНОЙ ФУНКЦИИ

Определение 2.1. Квадратичная функция есть сумма квадратичной формы $x^T Ax$, линейной формы $b^T x$ и константы:

$$f(x) = \frac{1}{2} x^T Ax + b^T x + c, \tag{2.1}$$

где A – симметричная матрица порядка n, $b \in R^n$, $c \in R^1$.

Определение 2.2. Квадратичная форма $x^T Ax$ и матрица A называются ***положительно определенными*** (ПО), если $x^T Ax > 0$ при любом $x \neq 0$. Если же $\forall x \;\; x^T Ax \geq 0$, то они ***положительно полуопределены*** (ППО).

Поведение квадратичной формы определяется собственными числами матрицы A.

Определение 2.3. Если выполнены условия $Ax = \lambda x$, где λ – число и $x \neq 0$, то столбец x называется ***собственным столбцом*** матрицы A. Скаляр λ называется ***собственным числом***.

Полный набор всех собственных чисел матрицы A будем обозначать через $Eig(A)$, наибольшее собственное число – через $EIG(A)$, наименьшее – через $eig(A)$. Собственные числа удовлетворяют уравнению $|A - \lambda E| = 0$, т. е. $Eig(A)$ есть набор из n корней характеристического полинома $|A - \lambda E|$ степени n, где E – единичная матрица размерности $n \times n$.

Теорема 2.1. Если матрица A – симметрична, то из ее собственных векторов можно построить ортонормированный базис $\mathcal{E} = (e_1, \ldots, e_n)$, $e_i \in E^n$, $i = \overline{1,n}$, в котором квадратичная форма имеет вид

$$x^T Ax = (y_1 \ldots y_n) \mathcal{E}^T A \mathcal{E} \begin{pmatrix} y_1 \\ \ldots \\ y_n \end{pmatrix} = (y_1 \ldots y_n) \operatorname{diag}[Eig(A)] \begin{pmatrix} y_1 \\ \ldots \\ y_n \end{pmatrix} = \sum_{i=1}^{n} \lambda_i y_i^2 . \tag{2.2}$$

Здесь y_i – координаты вектора x в ортонормированном базисе: $x = \mathcal{E}\, y$, где $y = (y_1 \ldots y_n)^T$.

Следствие 2.1. Из (2.2) следует, что для ПО матрицы A необходимо и достаточно положительности ее собственных чисел, а для ППО – их неотрицательности.

Если обозначить через m положительную оценку снизу для $eig(A)$ и через M – оценку сверху для $EIG(A)$, то имеет место следующая система неравенств, определяющая ПО квадратичной формы:

$$m\|x\|^2 \leq x^T Ax \leq M\|x\|^2, \qquad m > 0 . \tag{2.3}$$

Теорема 2.2. Пусть $f(x) \in C^2(R^n)$. Для того, чтобы функция $f(x)$ была выпуклой необходимо и достаточно, чтобы ее матрица вторых производных $f''(x)$ была ППО.

Следовательно, квадратичная функция (2.1), для которой выполняется (2.3), является выпуклой. Ее

производные вычисляются по формулам: $f'(x) = Ax + b$ и $f''(x) = A$. Поэтому имеет место следующая теорема.

Теорема 2.3. Если матрица A – ПО, то квадратичная функция (2.1) имеет в R^n единственную точку минимума x^*, удовлетворяющую системе линейных алгебраических уравнений (СЛАУ) $Ax^* + b = 0$.

3. МЕТОДЫ СПУСКА → ГРАДИЕНТНЫЕ МЕТОДЫ → МЕТОД НАИСКОРЕЙШЕГО ГРАДИЕНТНОГО СПУСКА

Определение 3.1. ***Методами спуска*** называется итерационные методы, применяемые для решения задачи минимизации функций нескольких переменных, для которых каждая итерация (шаг) приводит к уменьшению значения целевой функции

$$f\left(x^{k+1}\right) < f\left(x^k\right) \quad \forall \; k \ge 0. \tag{3.1}$$

В численных методах индекс итерации принято размещать сверху справа от обозначения итеративной точки, если она – векторная величина. Это позволяет сохранить обозначения компонент вектора с помощью нижнего индекса.

3.1. Общий план методов спуска

1°. *Выбирается начальное приближение – некоторая точка* x^0. Целесообразно использовать всю имеющуюся информацию о поведении целевой функции $f(x)$, чтобы выбрать x^0 поближе к точке минимума.

2°. Пусть приближение x^k к точке минимума найдено и $x^k \ne x^*$. *Выбирается* ***направление спуска****, т. е. вектор* $p^k \ne 0$*, такой, что для всех достаточно малых* $\alpha > 0$ *справедливо неравенство*

$$f\left(x^k + \alpha p^k\right) < f\left(x^k\right) \text{ (см. п. 3.2).}$$

3°. *Определяется величина* ***шага спуска по направлению*** p^k*, т. е. положительное число* $\alpha_k > 0$*, для которого выполняется неравенство*

$$f\left(x^k + \alpha_k p^k\right) < f\left(x^k\right) \text{ (см. п. 3.2).}$$

4°. *За очередное приближение к точке минимума принимается*

$$x^{k+1} = x^k + \alpha_k p^k. \tag{3.2}$$

5°. *Проверяется выполнение критерия окончания итераций* (см. п. 3.4 для $i = k + 1$). Если критерий выполняется, то итерации прекращаем и полагаем $x^* \approx x^{k+1}$. В противном случае возвращаемся к п. 2°.

Определение 3.2. Последовательность точек $\left\{x^k\right\}_{k=0}^{\infty}$, генерируемую методом спуска, называют ***траекторией спуска***.

3.2. Выбор направления и шага спуска. Метод наискорейшего спуска

Способ выбора направления спуска p^k определяет конкретный численный метод, а различные алгоритмы вычисления шага спуска α_k задают варианты этого метода.

Рассмотрим, каким образом выбирается направление p^k, которое обеспечивает убывание целевой функции на бесконечно малом положительном шаге α. Предполагая непрерывность первых частных производных целевой функции, разложим ее в ряд Тэйлора в точке x^k: $f\left(x^k + \alpha p^k\right) = f\left(x^k\right) + \alpha\left(f'\left(x^k\right), p^k\right) + o(\alpha)$, где $(\bullet, \bullet)$ – скалярное произведение. Тогда, если

$$\left(f'\left(x^k\right), p^k\right) < 0, \tag{3.3}$$

то $f\left(x^k + \alpha p^k\right) - f\left(x^k\right) < 0$. Таким образом, чтобы p^k было направлением спуска, необходимо и достаточно, чтобы p^k составляло острый угол с антиградиентом.

Определение 3.3. Методы, основанные на выборе в качестве направления спуска p^k антиградиента g^k или $-\beta f'\left(x^k\right)$, $\beta > 0$, называются ***градиентными***.

Градиент меняется при переходе из точки в точку, следовательно, *меняется и направление наискорейшего убывания функции*. Антиградиентное направление может не являться наилучшим из возможных направлений, поскольку всегда существует направление из текущей точки непосредственно в точку минимума (см. рис. 1.3), но мы его не знаем.

Будем рассматривать следующий способ выбора шага. Введем функцию одной скалярной переменной $F_k(\alpha) = f\left(x^k + \alpha\, p^k\right)$, и определим α_k, решая задачу одномерной минимизации

$$F_k(\alpha) \to \min_{\alpha \ge 0} \tag{3.4}$$

аналитически, например, для квадратичной функции, либо каким-либо итеративным методом.

Градиентный метод, в котором на каждой итерации используется шаг, являющийся решением задачи (3.4), был предложен в 1845 г. О. Коши и называется ***методом наискорейшего спуска***.

Отметим, что минимум по направлению не является минимумом целевой функции, если направление не указывает на точку минимума.

3.3. Метод наискорейшего градиентного спуска

Опишем алгоритм метода наискорейшего градиентного спуска (МНГС) для квадратичной функции (2.1). Пусть построена точка $x^k \in R^n$.

1°. Вычислим $g^k = -\left(Ax^k + b\right) \in R^n$ – градиентное направление спуска.

2°. Определим $\alpha_k \in R^1$ – шаг метода в данном направлении, из условия (3.4):

$$F_k(\alpha) = f\left(x^k + \alpha g^k\right) = f\left(x^k\right) - \alpha\left(g^k,\, g^k\right) + \frac{1}{2}\alpha^2\left(Ag^k,\, g^k\right).$$

В силу выпуклости квадратного трехчлена $F_k(\alpha)$ с положительным старшим коэффициентом, единственная точка минимума $\alpha_k \geq 0$, при условии, что $g^k \neq 0$, может быть найдена из необходимого условия экстремума $F_k'(\alpha) = -\left(g^k,\, g^k\right) + \alpha\left(Ag^k,\, g^k\right) = 0$, откуда получим:

$$\alpha_k = \frac{\left(g^k,\, g^k\right)}{\left(Ag^k,\, g^k\right)} > 0\,. \tag{3.5}$$

Очевидно, что при подстановке α_k в (3.2) построенная точка x^{k+1} удовлетворяет условию (3.1).

3°. Вычисляем следующую итерацию x^{k+1} по формуле (3.2). Проверяется выполнение критерия окончания итераций (см. (3.6)-(3.8) для $i = k+1$). Если критерий выполняется, то итерации прекращаем и полагаем $x^* \approx x^{k+1}$. В противном случае возвращаемся к п. 1°. Продолжая указанные построения, получим последовательность $\left\{x^k\right\}$ убывания функции $f(x)$.

3.4. Критерий окончания итераций

Критерием прекращения процесса вычислений на i-м шаге часто выбирается близость градиента к нулю

$$\left\|f'\left(x^i\right)\right\| < \varepsilon\,. \tag{3.6}$$

Помимо критерия (3.6), используются также:

$$\left\|x^i - x^{i-1}\right\| < \varepsilon_1, \text{ и } \left|f\left(x^i\right) - f\left(x^{i-1}\right)\right| < \varepsilon_2, \tag{3.7)-(3.8}$$

где $\varepsilon,\, \varepsilon_1, \varepsilon_2$ — заданные положительные числа. Нередко используют различные их сочетания, например, требуют, чтобы одновременно выполнялись условия (3.7) и (3.8) или даже все три условия (3.6)-(3.8).

3.5. Геометрический смысл МНГС

Из начальной точки x^0 стартует подвижная точка перпендикулярно линии уровня $\Gamma_0 = \left\{x \mid f(x) = f\left(x^0\right)\right\}$ в направлении антиградиента g^0 и движется до тех пор, пока не будет достигнуто минимальное на луче $L_0 = \left\{x \mid x = x^0 + \alpha g^0,\ \alpha \geq 0\right\}$ значение функции f. Из условия (3.4) получаем $F_0'(\alpha_0) = \left(f'\left(x^0 + \alpha_0 g^0\right),\, g^0\right) = \left(g^1,\, g^0\right) = 0$ при $\alpha_0 > 0$. Это равенство означает, что направление g^0 и, следовательно, луч L_0 являются касательными к линии уровня $\Gamma_1 = \left\{x \mid f(x) = f\left(x^1\right)\right\}$ в точке x^1. Из точки x^1 проводят спуск в перпендикулярном линии уровня Γ_1 направлении g^1, пока соответствующий луч L_1 не коснется в точке x^2 проходящей через эту точку линии уровня, и т. д. (См. рис. 3.1 а)).

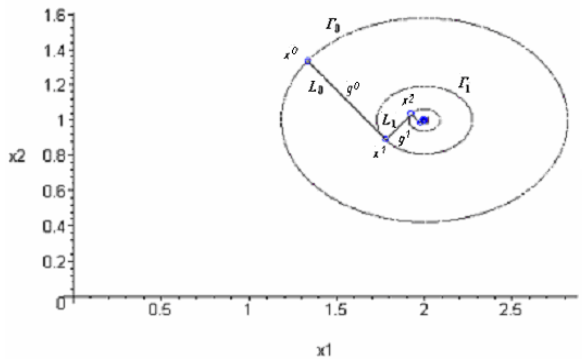

Рис. 3.3 а).

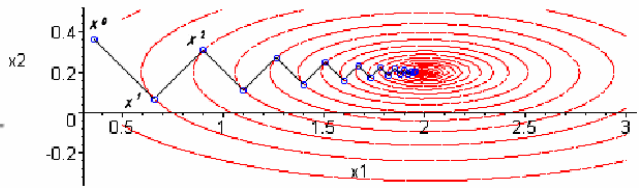

Рис. 3.3 б).

В градиентных методах спуск происходит по нормали к линии уровня, траектория спуска носит зигзагообразный характер и градиенты в любых двух последовательных точках ортогональны.

3.6. Сходимость и проблема "оврагов"

Градиентный метод сходится достаточно быстро, если для минимизируемой функции $f(x)$

поверхности уровня близки к сферам (при $n = 2$ – к окружностям, см. рис. 3.1 а)). Если же линии уровня сильно вытянуты в каком-то направлении, то по нормали к этому направлению целевая функция меняется значительно быстрее, чем вдоль направления. Такой характер целевой функции называется овражным (рис. 3.1 б)). Дно оврага к тому же может быть искривлено (рис. 3.2).

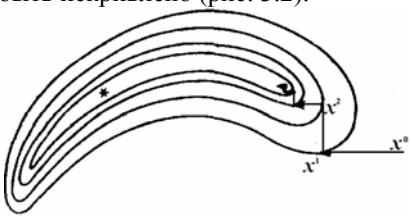

Рис. 3.2.

Если начальная точка x^0 попадает на "склон оврага", то направление градиентного спуска из этой и последующих точек будет почти перпендикулярным ко "дну оврага", хотя надо бы двигаться вдоль "дна оврага". Из-за малых шагов процесс практически останавливается вдали от точки минимума x^*, поэтому сходимость градиентного метода может быть очень медленной.

Поверхностями уровня квадратичной функции (в случае ПО) служат n-мерные эллипсоиды с центром в точке x^*. В двумерном случае графику ПО квадратичной формы соответствует поверхность, которая называется эллиптический параболоид. Линии уровня этой поверхности – концентрические эллипсы. Каноническое уравнение эллипса $x_1^2/a_1^2 + x_2^2/a_2^2 = 1$, где a_1 и a_2 его полуоси. Перепишем выражение (2.2) в виде

$$x^T Ax = \sum_{i=1}^{n} y_i^2 / (1/\lambda_i).$$

Очевидно, что длины полуосей эллипсоида (эллипса) обратно пропорциональны квадратным корням из собственных чисел. Если одно собственное число много меньше всех остальных, то ему соответствует полуось, длина которой много больше длин других полуосей, т.е. эллипсоид (эллипс) будет иметь сильно вытянутую форму. Отношение большей оси каждого из этих эллипсоидов (эллипсов) к меньшей оси равно $\nu(A) = EIG(A)/eig(A)$ – числу обусловленности матрицы A. Очевиден вывод: чем отношение $EIG(A)/eig(A)$ больше, тем ярче выражен овражный характер квадратичной функции, тем медленнее сходимость метода.

Теорема 3.1. Пусть A — симметричная ПО матрица и минимизируется квадратичная функция (2.1). Тогда при любом выборе начального приближения МНГС сходится к точке x^* минимума функции (2.1) со скоростью геометрической прогрессии знаменатель которой

$$q = (EIG(A) - eig(A))/(EIG(A) + eig(A)).$$

Причем если $eig(A)$ и $EIG(A)$ близки, то $EIG(A)/eig(A) \approx 1$ и q мало, то метод сходится достаточно быстро. Если же $EIG(A) >> eig(A)$, то $q \approx 1$, и следует ожидать медленной сходимости метода.

Теорема 3.3. Оценка расстояния от произвольной точки $x^k \in R^n$ до точки минимума x^* функции $f(x)$ может быть представлена в виде:

$$\left\| x^k - x^* \right\| \le \frac{1}{m} \left\| Ax^k + b \right\|. \tag{3.9}$$

3.7. Примеры

Пример 3.1. Решить задачу минимизации квадратичной функции $f(x_1, x_2) = x_1^2 + 2x_2^2 - 4x_1 - 4x_2$ с помощью метода наискорейшего спуска.

Запишем функцию в виде (2.1), где матрица $A = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$ симметричная ($A = A^T$) и ПО ($\lambda_1 = 2,\ \lambda_2 = 4$) и вектор $b = (-4, -4)^T$, и заметим, что $f(x_1, x_2) = (x_1 - 2)^2 + 2(x_2 - 1)^2 - 6$, поэтому точка минимума $x^* = (2,1)^T$ заранее известна. Пусть $x^0 = (0,0)^T$ – начальное приближение. Тогда

$$g^0 = -(Ax^0 + b) = (4,4)^T,\ \ \alpha_0 = \frac{(g^0,\ g^0)}{(Ag^0,\ g^0)} = 1/3,\ \ x^1 = x^0 + \alpha_0 g^0 = (4/3, 4/3)^T,\ \ \left\| Ax^1 + b \right\| = \sqrt{32}/3;$$

$$g^1 = -(Ax^1 + b) = (4/3, -4/3)^T,\ \ \alpha_1 = 1/3,\ \ x^2 = x^1 + \alpha_1 g^1 = (16/9,\ 8/9)^T,\ \ \left\| Ax^2 + b \right\| = \sqrt{32}/3^2;$$

$$g^2 = -(Ax^2 + b) = (4/9, 4/9)^T,\ \ \alpha_1 = 1/3,\ \ x^2 = x^1 + \alpha_1 g^1 = (52/27,\ 28/27)^T,\ \ \left\| Ax^2 + b \right\| = \sqrt{32}/3^3.$$

Заметим, что $\left\| Ax^k + b \right\| = \sqrt{32}/3^k \xrightarrow[k\to\infty]{} 0$,. Таким образом, последовательность x^k, полученная по

МНГС, сходится со скоростью геометрической прогрессии, знаменатель которой $q = 1/3$. Траектория спуска представлена на рис. 3.3 а).

Пример 3.2. Применение МНГС для минимизации квадратичной функции $f(x_1, x_2) = x_1^2 + 10x_2^2 - 4x_1 - 4x_2$ с симметричной и ПО матрицей $A = \begin{pmatrix} 2 & 0 \\ 0 & 20 \end{pmatrix}$ ($\lambda_1 = 2,\ \lambda_2 = 20$) при $x^0 = (0,0)^T$ дает последовательность приближений

$$g^0 = -(Ax^0 + b) = (4,4)^T,\quad \alpha_0 = 1/11,\ x^1 = x^0 + \alpha_0 g^0 = (4/11, 4/11)^T,\ \|Ax^1 + b\| = 2^2 3^2 \sqrt{2}/11;$$

$$g^1 = -(Ax^1 + b) = (36/11, -36/11)^T,\ \alpha_1 = 1/11,\ x^2 = x^1 + \alpha_1 g^1 = (80/11^2,\ 8/11^2)^T,$$

$$\|Ax^2 + b\| = 2^2 3^4 \sqrt{2}/11^2;\quad g^2 = -(Ax^2 + b) = (324/11^3, 324/11^3)^T,\ \alpha_1 = 1/11,$$

$$x^2 = x^1 + \alpha_1 g^1 = (1204/11^3,\ 412/11^3)^T,\quad \|Ax^2 + b\| = 2^2 3^6 \sqrt{2}/11^3;$$

$\|Ax^k + b\| = 4\sqrt{2}(9/11)^k \xrightarrow[k\to\infty]{} 0$, где $x^* = (2,\ 0.2)^T$. Траектория спуска изображена на рис. 3.3 б). Последовательность x^k сходится здесь со скоростью геометрической прогрессии, знаменатель которой равен $q = 9/11$, т. е. существенно медленнее, чем в предыдущем примере.

4. МЕТОД ПОКООРДИНАТНОГО СПУСКА

В этом методе в качестве очередного направления спуска выбирается направление одной из координатных осей $p^k = e^i = (0,\ldots,0,1_i,0,\ldots,0)^T$, а шаг спуска по формуле (3.4).

4.1. Геометрическая интерпретация

Геометрический смысл метода состоит в движении в направлениях, параллельных координатным осям (см. рис. 4.1 а)): точка (x_0, y_0) описывает начальное приближение. Проводя спуск по координате x, попадем в точку (x_1, y_0). Двигаясь параллельно оси ординат, придем в точку (x_1, y_1) и т. д. При движении вдоль координатной оси идем до точки минимума, т.е. касания некоторой линии уровня. Число шагов зависит от начального приближения и, главное, от ориентации линий уровня относительно координатных осей. Так, в задаче минимизации квадратичной функции при ориентации осей эллипсов вдоль координатных осей достаточно однократного применения цикла. В овражных случаях при произвольной ориентации относительно координатных осей даже в квадратичной задаче метод работает плохо (см. рис. 4.1 б)).

Рис. 4.1 а).

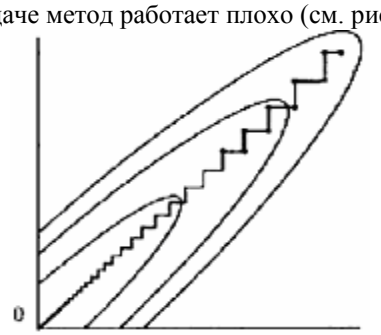

Рис. 4.1 б).

4.2. Примеры

Пример 4.1. Рассмотрим функцию $f(x) = x_1^2 + x_1 x_2 + x_2^2$, в матричном виде $f(x) = \frac{1}{2} x^T A x$, где $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Собственные числа матрицы A: $\lambda_1 = 3;\ \lambda_2 = 1$. Построим собственные столбцы матрицы A.

$1^\circ\ \lambda_1 = 3;\ \Rightarrow\ \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} x = 0;\ \Rightarrow\ -x_1 + x_2 = 0;\ \Rightarrow\ X = \begin{pmatrix} 1 \\ 1 \end{pmatrix} C,\ C \in E^1,\ \|X\| = \sqrt{2}$.

$2^\circ.\ \lambda_2 = 1;\ \Rightarrow\ \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} x = 0;\ \Rightarrow\ x_1 + x_2 = 0;\ \Rightarrow\ X = \begin{pmatrix} 1 \\ -1 \end{pmatrix} C,\ C \in E^1,\ \|X\| = \sqrt{2}$.

Ортонормированные собственные векторы $e_1 = (1/\sqrt{2}, 1/\sqrt{2})^T$ и $e_2 = (1/\sqrt{2}, -1/\sqrt{2})^T$, соответственно. Так как $y = \mathcal{E}^{-1} x$, а $\mathcal{E}^{-1} = \mathcal{E} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$, то

$y = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2}\,(x_1 + x_2) \\ 1/\sqrt{2}\,(x_1 - x_2) \end{pmatrix}$. Таким образом, новые переменные $y_1 = 1/\sqrt{2}\,(x_1 + x_2)$ и $y_2 = 1/\sqrt{2}\,(x_1 - x_2)$ и $F(x) = \frac{1}{2}\left(3y_1^2 + y_2^2\right) = \frac{3}{4}(x_1 + x_2)^2 + \frac{1}{4}(x_1 - x_2)^2$.

Линии уровня целевой функции — эллипсы с центром в начале координат, так как отсутствует линейная часть (рис. 4.2).

1°. Пусть задана $x^0 = (1, 1)^T$, $f(x^0) = 3$. Фиксируем $x_2 = 1$ и решаем задачу $\frac{3}{4}(x_1 + 1)^2 + \frac{1}{4}(x_1 - 1)^2 \to \min\limits_{x_1}$. Минимум достигается в точке $x_1 = -1/2$ и равен $3/4$.

2°. Фиксируем $x_1 = -1/2$ и ищем минимум функции $\frac{3}{4}\left(x_2 - \frac{1}{2}\right)^2 + \frac{1}{4}\left(x_2 + \frac{1}{2}\right)^2$, который достигается в точке $x_2 = 1/4$ и равен $3/16$. Получаем точку $(-1/2, 1/4)^T$.

3°. Фиксируем $x_2 = 1/4$ и находим $x_1 = -1/8$.

4°. Фиксируем $x_1 = -1/8$ и находим $x_2 = 1/16$. В точке $(-1/8, 1/16)$ значение целевой функции $3/256$. Так как $\left|f(-1/8, 1/16) - f(-1/2, 1/4)\right| = 3/256 < 0.05$, то процесс завершен.

МНГС из той же начальной точки $(1, 1)^T$ дает минимум за одну итерацию, так как $g^0 = -\left(Ax^0 + b\right)^T = -(2x_1 + x_2,\ x_1 + 2x_2)^T = (-3, -3)^T$ и при $\alpha_0 = \left(g^0, g^0\right)/\left(Ag^0, g^0\right) = 1/3$ мы получаем точку минимума $x^1 = x^0 + \alpha_0 g^0 = (0, 0)^T$.

С другой стороны, если взять за начальное приближение точку $\left(0, \sqrt{3}\right)^T$, которая лежит на той же линии уровня, что и точка $(1, 1)^T$, и первой изменять x_2, а не x_1, то ***покоординатный спуск*** дает ***минимум за один шаг***, в то время как ***наискорейший спуск*** $g^0 = \left(-\sqrt{3}, -2\sqrt{3}\right)^T$ из точки $\left(0, \sqrt{3}\right)^T$ дает за одну итерацию точку $x^1 = \left(-5\sqrt{3}/14,\ 4\sqrt{3}/14\right)^T$, в которой целевая функция имеет значение $9/28$. Если сделать еще один шаг по антиградиенту $g^1 = \left(6\sqrt{3}/14,\ -3\sqrt{3}/14\right)^T$, то получим (при $\alpha_1 = 5/6$) точку $\left(0,\ 3\sqrt{3}/28\right)^T$, значение целевой функции в ней $27/784$. Таким образом, ***после двух итераций*** по одной из координат отклонение от точки минимума составляет примерно 0.17, а по целевой функции — примерно 0.034.

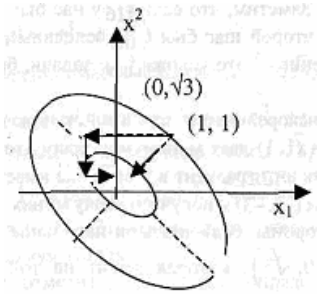

Рис. 4.2.

5. МЕТОД СОПРЯЖЕННЫХ ГРАДИЕНТОВ

5.1. Алгоритм метода

Идея этого метода в том, чтобы на каждом шаге в качестве направления спуска использовать не антиградиент, а его линейную комбинацию с прежним направлением спуска. Последовательность векторов спуска p^k строится следующим образом:

1. $p^0 = g^0$, т.е. первый шаг делаем по антиградиенту,
2. $p^{k+1} = g^{k+1} + \beta_k p^k$, где $\beta_k = \left\|g^{k+1}\right\| / \left\|g^k\right\|$.
3. $x^{k+1} = x^k + \alpha_k p^k, \quad k = 0, 1, 2, \ldots$.

Заметим, что при вычислении β_k используются только градиенты, а для вычисления нового направления нужно помнить и использовать старое направление, а не старый антиградиент. В методе сопряженных градиентов новое и старое направления не ортогональны. В случае минимизации положительно определенной квадратичной формы с матрицей A все направления спуска p^i оказываются

A-ортогональными, т.е. удовлетворяют условию $\left(p^i, Ap^j\right)=0$ при любом $i \neq j$. Такие ***векторы называются сопряженными***, из чего и происходит название метода.

При минимизации квадратичной формы с матрицей A, независимо от того, как получено направление спуска p^i, для α_k справедлива формула, аналогичная формуле (3.6) для МНГС,

$$\alpha_k = \frac{\left(p^k, g^k\right)}{\left(Ap^k, p^k\right)}. \tag{5.1}$$

Теорема 5.1. Если целевая функция от n переменных представляет собой квадратичную форму, то алгоритм сопряженных градиентов дает точку минимума не более чем за n шагов.

Практически же в овражных случаях из-за неточности вычисления шага может потребоваться несколько больше шагов, в некоторых случаях метод может зацикливаться.

5.2. Геометрическая интерпретация

Сделав из начальной точки x^0 шаг по методу наискорейшего спуска, вычислим антиградиент g^1 в новой точке x^1. Будем использовать вместо антиградиента g^1 направление спуска $p^1 = g^1 + \beta_0 g^0$, в направлении которого нужно дойти до точки минимума x^2, в ней вычислить антиградиент g^2, $\beta_1 = \left\|g^2\right\| / \left\|g^1\right\|$ и $p^2 = g^2 + \beta_1 p^1$ и т.д.

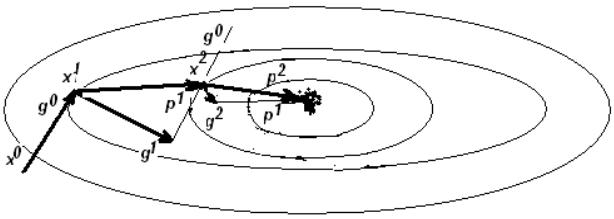

Рис. 5.1.

5.3. Примеры

Пример 5.1. Рассмотрим метод сопряженных градиентов на примере квадратичной функции из примера 4.1 $f(x) = x_1^2 + x_1 x_2 + x_2^2$. Пусть $x^0 = \left(0, \sqrt{3}\right)^T$, $p^0 = g^0 = \left(-\sqrt{3}, -2\sqrt{3}\right)^T$, $x^1 = \left(-5\sqrt{3}/14,\ 4\sqrt{3}/14\right)^T$, $g^1 = \left(6\sqrt{3}/14,\ -3\sqrt{3}/14\right)^T$.

Сделаем второй шаг не по антиградиенту, а по сопряженному направлению p^1. Вычислим $\beta_0 = \left\|g^1\right\| / \left\|g^0\right\| = \dfrac{9}{14^2}$. Далее $p^1 = g^1 + \beta_0 g^0 = \left(75\sqrt{3}/14^2, -60\sqrt{3}/14^2\right)^T$. Точки на луче $x^1 + \alpha p^1$ имеют координаты $\sqrt{3}/14\left(-5 + \alpha_2\, 75/14\right)$ и $\sqrt{3}/14\left(4 - \alpha_2\, 60/14\right)$. При $\alpha_2 = 14/15$ обе координаты обращаются в нуль.

Согласно теореме 5.1, метод сопряженных градиентов при минимизации квадратичной функции дает точку минимума не более чем за n шагов (у нас $n = 2$) независимо от выбора начальной точки. Это означает, что направление p^1 ведет в точку минимума $(0,\ 0)^T$.

6. ТИПОВОЙ РАСЧЕТ

1. Программная реализация на языке $C++$ метода решения задачи минимизации (2.1). Метод определяется преподавателем. Построить линии уровня с указанием векторов спуска.
2. Найти решение x^* СЛАУ $Ax + b = 0$. Вывести точное x^* и приближенное x^k решение.
3. Сравнить $\left\|x^* - x^k\right\|$ с $\left\|x^k - x^{k-1}\right\|$. Проверить вычисления при различных начальных векторах x^0 и проследить за зависимостью k от x^0. Указать скорость сходимости q.

Параметр N равен номеру студента по списку группы.

$$A = \begin{pmatrix} N & -N \\ -N & N+1 \end{pmatrix}; \quad b = \begin{pmatrix} N-3 \\ N+5 \end{pmatrix}.$$

ЛИТЕРАТУРА

1. Васильев Ф. П. Численные методы решения экстремальных задач, М. 1988.
2. Пшеничный Б.Н., Данилин Ю. М. Численные методы в экстремальных задачах, М. 1975.
3. Карманов В. Г. Математическое программирование, М., 1986.