Методы минимизации функций при отсутствии ограничений: обзор

Пример: задача идентификации – требуется найти вектор θ из условия минимума критерия

$$J_N(\theta) = 1/(2N)\sum_{k=0}^{N} \| y(k+1) - g(k, \varphi(k), \theta) \|^2.$$

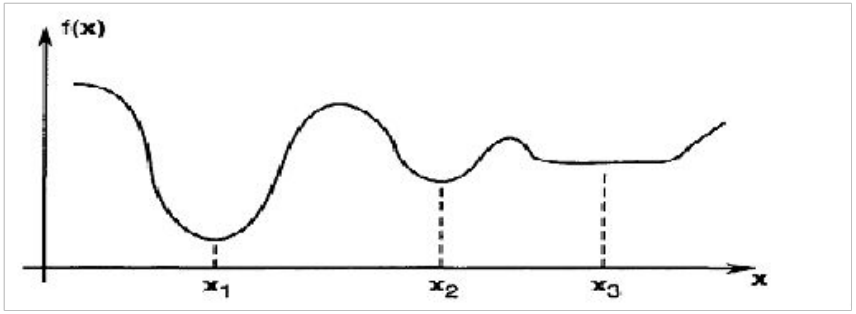
1 Условия минимума гладких функций

Пусть задана функция $f(x)$: $\mathrm{R}^n \to R$.

Определение 1. Точка x^* называется локальным минимумом $f(x)$ на R^n, если найдется $\varepsilon > 0$ такое, что $f(x) \geq f(x^*)$ для всех x из ε-окрестности x^*, т.е., $\| x - x^* \| \leq \varepsilon$, где $\| x \|^2 = (x_1^2 + ... + x_n^2)$– евклидова норма.

Определение 2. Точка x^* называется глобальным минимумом $f(x)$ на R^n, если $f(x) \geq f(x^*)$ для всех $x \in R^n$.

Пример 1. x_1 - глобальный минимум, x_2, x_3 - локальные минимумы.



Пример 2. Отсутствие минимума.

$f(x) = e^{-x}$

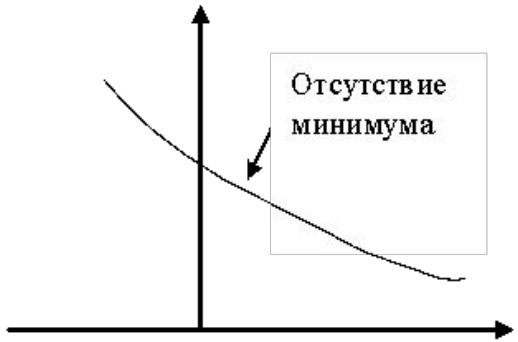


Рис.2

Теорема 1. (Ферма). Пусть x^* точка минимума $f(x)$ на R^n и $f(x)$ дифференцируема в точке x^*. Тогда

$$\nabla f(x^*) = 0, \qquad (1)$$

где $\nabla f(x) = (\partial f(x)/\partial x_1, ..., \partial f(x)/\partial x_n)^T$ - градиент $f(x)$.

Теорема 2. Пусть в точке x^* $f(x)$ дважды дифференцируема и выполняется необходимое условие

$$\nabla f(x^*) = 0$$

и при этом

$$\nabla^2 f(x^*) > 0, \qquad (2)$$

где

$$\nabla^2 f(x) = \begin{pmatrix} \partial^2 f(x)/\partial x_1^2 & . & . & . & \partial^2 f(x)/\partial x_1 \partial x_n \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ \partial^2 f(x)/\partial x_1 \partial x_n & . & . & . & \partial^2 f(x)/\partial x_n^2 \end{pmatrix}$$ - Гессиан.

Тогда x^* - точка локального минимума.

Отметим, что Гессиан – симметричная матрица, а условие $A = \nabla^2 f(x^*) > 0$ означает положительную определенность этой матрицы. Т.е., для $\forall y \in R^n, y \neq 0$ выполняется условие

$$y^T A y > 0.$$

Опираясь на эти утверждения, задача поиска минимума может быть сведена к поиску решений системы нелинейных уравнений

$$\nabla f(x) = 0$$

и проверке выполнения условия (2) на полученном решении.

2 Градиентный метод

Пусть $f(x) \in C^1(R^n)$. В этом случае наиболее простым методом минимизации $f(x)$ является градиентный

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),\ k = 0,1,..., \qquad (3)$$

где x_0 - задано, α_k - определяет длину шага.

Основан на следующей идее. Мы начинаем в некоторой точке x_0 и генерируем векторы $x_1, x_2, ...$ такие, что $f(x)$ уменьшается в каждой итерации, т.е.

$$f(x_{k+1}) < f(x_k). \qquad (4)$$

Имеем

$$f(x_{k+1}) = f(x_k) + \nabla f^T(x_k)(x_{k+1} - x_k) + o(\| x_{k+1} - x_k \|) =$$

$$= f(x_k) - \alpha_k \| \nabla f(x_k) \|^2 + o(\| x_{k+1} - x_k \|),$$

где $\varphi(z) = o(\| z \|)$ - функция, удовлетворяющая условию

$$\lim_{z \leftarrow 0} \varphi(z)/z = 0.$$

При α_k достаточно малом второе слагаемое доминирует над третьим и $f(x_{k+1}) < f(x_k)$.

Первая проблема, возникающая при реализации метода это выбор α_k. Достаточно малый шаг обеспечит убывание функции, но может привести к неприемлемо большому количеству итераций, необходимого для достижения точки минимума. С другой стороны, слишком большой шаг может приводить к невыполнению условия (4), либо к колебательности около точки минимума.

Пример. $f(x) = ax^2, a > 0$.

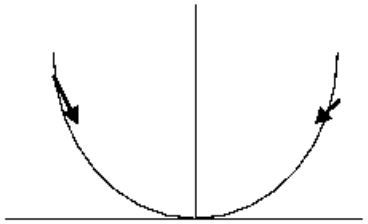


Рис.2

Алгоритм принимает вид

$$x_{k+1} = (1 - 2\alpha_k a)x_k .$$

Пусть $\alpha_k = const$ и $0 < \alpha < 1/a$. Тогда последовательность будет сходиться к $x^* = 0$, а при $\alpha > 1/a$ расходиться.

Один из возможных вариантов выбора α_k состоит в обеспечении максимального уменьшения $f(x)$ на каждой итерации

$$\alpha_k = \arg \min_{\alpha>0} f(x_k - \alpha \nabla f(x_k)) . \quad (5)$$

В этом случае градиентный метод (ГМ) называются методом наискорейшего спуска (МНС).

Если для некоторого k

$$\nabla f(x_k) = 0 ,$$

то точка x_k удовлетворяет условию (1). Мы можем использовать в качестве основы для правила остановки алгоритма:

$$\| \nabla f(x_k) \| < \varepsilon , \quad (6)$$

где $\varepsilon > 0$ - заданное пороговое значение.

Возможные альтернативы:

$$| f(x_{k+1}) - f(x_k) | < \varepsilon , \quad (7)$$

$$\| x_{k+1} - x_k \| < \varepsilon . \quad (8)$$

Или, используя относительные величины (аналоги (6)-(8))

$$| f(x_{k+1}) - f(x_k) | / | f(x_k) | < \varepsilon , \quad (9)$$

$$\| x_{k+1} - x_k \| / \| x_k \| < \varepsilon . \quad (10)$$

Относительные критерии более предпочтительны, так как не зависят от шкалы измерения переменных. Для того, чтобы избежать деления на очень малые числа, критерии останова можно модифицировать:

$$| f(x_{k+1}) - f(x_k) | / \max(1, | f(x_k) |) < \varepsilon . \quad (11)$$

ГМ является примером итерационного алгоритма. Это означает, что алгоритм генерирует последовательность точек, каждую на базе предыдущей.

Определение 3. Говорят, что итерационный алгоритм глобально сходится, если для любой начальной точки алгоритм генерирует последовательность точек, сходящуюся к точке, удовлетворяющей необходимому условию (1).

Это можно сформулировать еще так: для $\forall x_0 \in R^n$, $\forall \varepsilon > 0$, $\exists N > 0$ такие, что при $k > N$ будет выполняться условие $\| x_k - x^* \| < \varepsilon$.

В случае, когда алгоритм глобально не сходится , он может все еще генерировать сходящую последовательность при условии, что начальная точка достаточно близка к точке минимума. Насколько близко к решению должна быть начальная точка зависит от алгоритма.

Формальное определение локальной сходимости: для, $\forall \varepsilon > 0$, $\exists \delta > 0$ $\exists N > 0$ такие, что при $\| x_0 - x^* \| < \delta$ будет выполняться условие $\| x_k - x^* \| < \varepsilon$ для всех $k > N$.

Со сходимостью алгоритма тесно связана скорость его сходимости, т.е. насколько быстро алгоритм сходится к решению. Свойства ГМ обсудим на примере квадратичной функции

$$f(x) = 1/2 x^T Q x - b^T x, \qquad (12)$$

где $b \in R^n$- постоянный вектор, Q - положительно определенная, симметричная матрица ($Q > 0$).

Единственная точка минимума определяется выражением

$$\nabla f(x) = Qx - b = 0.$$

Откуда

$$x^* = Q^{-1} b.$$

То, что это действительно точка минимума следует из выражения

$$\nabla^2 f(x^*) = Q > 0.$$

Мы воспользовались следующими соотношениями при вычислении градиента:

$$\frac{\partial}{\partial x_i} b^T x = b^T \frac{\partial}{\partial x_i} x = b^T e_i,$$

$$\frac{\partial}{\partial x_i} x^T Q x = (\frac{\partial}{\partial x_i} x)^T Q x + x^T Q \frac{\partial}{\partial x_i} x = e_i^T Q x + x^T Q e_i = 2 e_i^T Q x ,$$

где e_i - i-единичный вектор.

Теорема 3. В МНС $x_k \to x^*$ для любого x_0.

Теорема 4. Для ГМ с фиксированным шагом $x_k \to x^*$ для любого x_0, если и только если

$$0 < \alpha < 2 / \lambda_{\max}(Q) ,$$

где $\lambda_{\max}(Q)$ - максимальное собственное число матрицы Q.

Собственные числа матрицы являются корнями ее характеристического уравнения

$$\det(\lambda I - Q) = \lambda^n + ... + q_n = 0 .$$

Пусть последовательность x_k сходится к x^*, т.е.

$$\lim_{k \to \infty} \| x_k - x^* \| = 0 .$$

Определение 4. Говорят, что порядок сходимости метода есть $p \in R$, если

$$\| x_{k+1} - x^* \| = O(\| x_k - x^* \|^p)$$

где $\varphi(z) = O(\| z \|)$ - функция, удовлетворяющая условию $| \varphi(z) | / \| z \| < const$ при $z \to 0$.

Или по другому $\| x_{k+1} - x^* \| \le C \| x_k - x^* \|^p$, при $\| x_k - x^* \| \to 0$

При $p = 1$ - линейная сходимость, $p = 2$ - квадратичная сходимость.

Теорема 5. Порядок сходимости МНС для квадратичной функции равен 1.

3 Метод Ньютона

ГМ использует только производные первого порядка при генерировании последовательности, сходящейся к решению. Использование производных более высокого порядка может приводить к более эффективным алгоритмам.

С помощью формулы Тейлора можно получить квадратичную аппроксимацию $f(x)$:

$$f(x) = f(x_k) + \nabla f^T(x_k)(x - x_k) + 1/2(x - x_k)^T \nabla f^2(x_k)(x - x_k).$$

Использование необходимого условия (1) дает

$$\nabla f(x_k) + \nabla f^2(x_k)(x - x_k) = 0.$$

Откуда следует метод Ньютона (МН)

$$x_{k+1} = x_k - (\nabla f^2(x_k))^{-1} \nabla f(x_k), \qquad (13)$$

в предположении, что $\nabla f^2(x_k)$ невырожденная матрица.

Условие (4) $f(x_{k+1}) < f(x_k)$ может не выполняться для МН, например, если x_0 далеко от решения. Несмотря на это МН быстрее сходится, если x_0 достаточно близко к решению.

Для квадратичной функции нетрудно показать, что метод сходится за один шаг. Действительно, пусть

$$f(x) = 1/2 x^T Q x - b^T x,$$

где $b \in R^n$ - постоянный вектор, Q - положительно определенная, симметричная матрица ($Q > 0$). Тогда

$$\nabla f(x) = Qx - b, \ \nabla^2 f(x) = Q.$$

Откуда следует

$$x_1 = x_0 - Q^{-1}(Qx_0 - b) = Q^{-1}b = x^*.$$

Общий случай.

Теорема 6. Пусть $f(x) \in C^2(R^n)$, $\nabla f(x^*) = 0$ и $\nabla^2 f(x^*) \neq 0$. Тогда для любого x_0 достаточно близкого к x^* МН сходится к x^* с порядком сходимости равным 2.

4 Метод Левенберга-Маркардта

Если матрица $\nabla f^2(x_k)$ не является положительно определенной, то последовательность, генерируемая МН может не удовлетворять условию $f(x_{k+1}) < f(x_k)$. Предложена модификация МН

$$x_{k=1} = x_k - (\nabla f^2(x_k) + \mu_k I)^{-1} \nabla f(x_k), \qquad (14)$$

где $\mu_k \geq 0$.

Если $\mu_k \to 0$, то имеем МН, а если $\mu_k \to \infty$, то - ГМ. При использовании метода начинают с большого μ_k, а затем медленно уменьшают его значение до тех пор пока выполнится условие $f(x_{k+1}) < f(x_k)$.

5 Метод Гаусса-Ньютона

Рассмотрим задачу минимизации функции

$$f(x) = 1/2 \sum_{i=1}^{m} r_i^2(x), \qquad (15)$$

где $r_i(x)$: $\mathrm{R}^n \to R$, $i = 1,...,m$ - заданные функции.

Пример. Пусть требуется найти минимум функции вида

$$J_N(\theta) = 1/(2N) \sum_{k=0}^{N} (y(k+1) - g(k, \varphi(k), \theta))^2.$$

Тогда $r_i(x) = y(i+1) - g(i, \varphi(i), x)$, $i = 1,2,...,N$.

Удобно перейти к векторно-матричным обозначениям

$$f(x) = 1/2 r^T(x) r(x),$$

где $r(x) = (r_1(x),..., r_m(x))^T$.

Для того, что бы воспользоваться МН необходимо найти градиент и Гессиан. Имеем

$$(\nabla f(x))_j = \partial f(x)/\partial x_j = \sum_{i=1}^{m} r_i\ (x)\partial r_i(x)/\partial x_j .$$

Введем матрицу первых производных от $r(x)$ (Якобиан)

$$J(x) = \begin{pmatrix} \partial r_1(x)/\partial x_1 & . & . & \partial r_1(x)/\partial x_n \\ . & & & . \\ . & & & . \\ \partial r_m(x)/\partial x_1 & . & . & \partial r_m(x)/\partial x_n \end{pmatrix}.$$

Градиент может быть представлен в виде

$$\nabla f(x) = J^T(x) r(x).$$

(k, j) элемент Гессиана определяется выражением

$$\partial^2 f(x)/\partial x_k \partial x_j = \partial/\partial x_k (\partial f(x)/\partial x_j) =$$

$$= \partial/\partial x_k (\sum_{i=1}^{m} r_i\ (x)\partial r_i(x)/\partial x_j) = \sum_{i=1}^{m} (\partial r_i\ (x)/\partial x_k \partial r_i(x)/\partial x_j + r_i(x)\partial^2 r_j(x)/\partial x_k \partial x_j).$$

Пусть $S(x)$ матрица (k, j) компоненты, которой определяются выражением

$$\sum_{i=1}^{m} r_i(x)\partial^2 r_j(x)/\partial x_k \partial x_j .$$

Тогда Гессиан может быть представлен в виде

$$\nabla^2 f(x) = J^T(x)J(x) + S(x).$$

Откуда следует представление для МН

$$x_{k=1} = x_k - (J^T(x)J(x) + S(x))^{-1} J^T(x) r(x).$$

В некоторых случаях матрицей $S(x)$ можно пренебречь в этом выражении и МН сводится к методу Ньютона-Гаусса (МНГ)

$$x_{k+1} = x_k - (J^T(x)J(x))^{-1} J^T(x)r(x). \qquad (16)$$

Матрица $J^T(x)J(x)$ может вырождаться. В этом случае можно использовать модификацию метода

$$x_{k+1} = x_k - (J^T(x)J(x) + \mu_k I)^{-1} J^T(x)r(x). \qquad (17)$$

Можно интерпретировать $\mu_k I$, как аппроксимацию $S(x)$.

5 Обсуждение

Задача. Пусть дана функция $f(x)\colon \mathrm{R}^n \to R$.

Требуется найти $\min f(x),\ \forall x \in R^n$.

Алгоритмы

1. ГМ

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),\ k = 0,1,..., \qquad (3)$$

где x_0 - задано, α_k - определяет длину шага.

2 НМ.

$$x_{k=1} = x_k - (\nabla f^2(x_k))^{-1} \nabla f(x_k), \qquad (13)$$

3. НГМ. Ищется минимум функции специального вида

$$f(x) = 1/2 \sum_{i=1}^{m} r_i^2(x), \qquad (15)$$

где $r_i(x)\colon \mathrm{R}^n \to R$, $i = 1,...,m$ - заданные функции.

$$x_{k=1} = x_k - (J^T(x)J(x))^{-1} J^T(x) r(x). \qquad (16)$$

4. МЛМ

$$x_{k=1} = x_k - (J^T(x)J(x) + \mu_k I)^{-1} J^T(x) r(x). \qquad (17)$$

Свойства алгоритмов

1. Все итерационные
2. МГ самый простой и медленно сходящийся
3. Могут "застревать" в локальных минимумах. Почему это плохо?

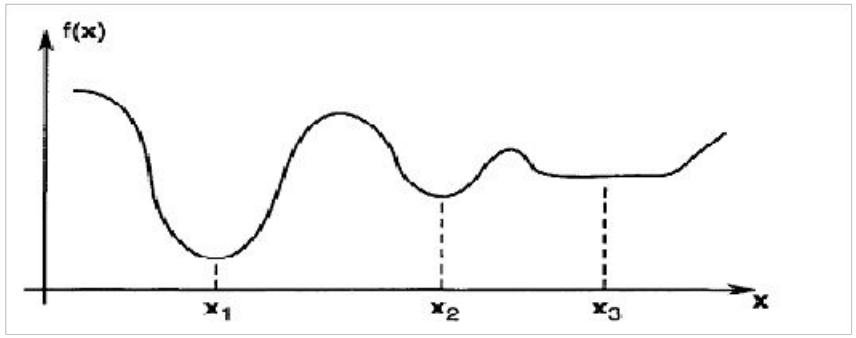


Рис.1

Пример. Пусть требуется аппроксимировать данные $(t, y(t))$, $t = 1,..., N$ функцией $y = g(t, x)$. Воспользуемся нелинейным методом наименьших квадратов

$$f(x) = 1/(2N)\sum_{k=0}^{N}(y(k) - g(k, x))^2 \to \min.$$

На рис.3 показаны две возможные аппроксимации.

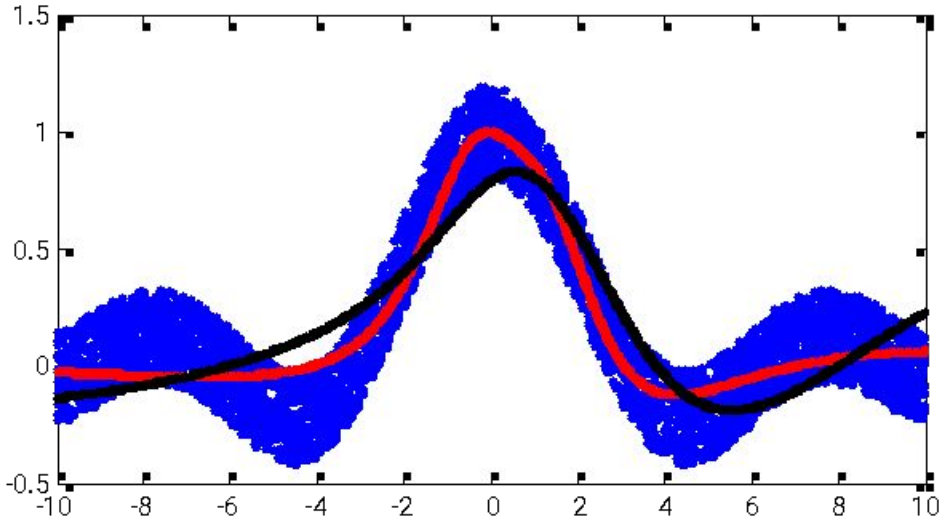


Рис.3

4. Самым эффективным с точки зрения количества итераций является МЛМ с изменяющимся параметром . Итерации начинаются с большого значения и затем оно меняется таким образом, чтобы обеспечить выполнение условия .