Методы поиска экстремума

Д.Е.Бурланков

Нижегородский университет им. Н.И. Лобачевского

2006

Стратегия поиска экстремума

Поиск экстремума – минимума или максимума всегда может быть сведен к поиску минимума: если исходно нужен максимум, смена знака функции приводит задачу к поиску минимума. Минимумов может быть много. Минимум, в котором функция имеет наименьшее значение среди всех минимумов, называется *глобальным минимумом*. Все остальные называются *локальными минимумами*.

Стратегия:

- Нахождение области экстремума.
- Нахождение экстремума внутри локализованной области. Уменьшение области поиска.
- Уменьшение шага поиска.

Одномерные задачи

Совершается обход области с постоянным шагом и изучается поведение функции

- Деление отрезка пополам
- Деление золотым сечением

Метод квадратичной аппроксимации

В трех точках: x_1, x_2, x_3 вычисляются значения функции y_1, y_2, y_3 и по этим трем точкам строится парабола (методом Лагранжа):

$$y(x) = y_1 \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} + y_2 \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} + y_3 \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2}$$

Ее экстремум

$$\frac{d\,y(x)}{d\,x} = y_1 \frac{(2\,x - x_2 - x_3)}{(x_1 - x_2)(x_1 - x_3)} + y_2 \frac{(2\,x - x_1 - x_3)}{(x_2 - x_1)(x_2 - x_3)} + y_3 \frac{(2\,x - x_1 - x}{(x_3 - x_1)(x_3 -}$$

определяет новую точку

$$x_4 = \frac{x_1^2(y_2 - y_3) + x_2^2(y_3 - y_1) + x_3^2(y_1 - y_2)}{2\,(x_1\,(y_2 - y_3) + x_2\,(y_3 - y_1) + x_3\,(y_1 - y_2))}$$

Однако, прежде чем вычислять эту точку, нужно вычислить знаменатель:

- Если он положителен – в точке x_4 минимум параболы (но не обязательно искомой функции).
- Если он отрицателен – в точке x_4 максимум параболы (но не обязательно искомой функции).
- Если он равен нулю – точки лежат на одной прямой.

Затем отбрасывается точка, наиболее удаленная от x_4 и на оставшихся трех точках алгоритм повторяется.

Многомерные экстремумы

Поиск экстремума имеет стратегию:

- Поиск области экстремума. В одномерном случае – шаги в направлении убывания функции, пока она не станет возрастать. Затем можно уменьшить шаг.
- Поиск в найденной области

Симплексный поиск

В n - мерном пространстве выбирается $n+1$ начальных точек, в которых вычисляется значение минимизируемой функции, и координаты точки, в которой значение функции максимально (дальше всего от минимума) инвертируются по формуле:

$$\bar{\mathbf{r}}_{n+1} = \mathbf{r}_c - (\mathbf{r}_{n+1} - \mathbf{r}_c) = 2\,\mathbf{r}_c - r_{n+1}; \quad \bar{x}^i_{n+1} = \frac{2}{n}\,(x^i_1 + x^i_2 + \cdots + x^i_n) - x^i_{n+1}.$$

В этой точке также вычисляется значение функции, и если оно опять оказывается максимальным, то минимум находится внутри области, составленной из исходного и инвертируемого симплексов.
В противном случае нужно инвертировать симплекс в вершине с максимальным значением функции. При этом максимальное значение все время уменьшается.

Поиск на кубической решетке

Более простая стратегия переноса одномерного поиска на многомерный – разбиение области n - мерного пространства на n - мерные кубы – построение n - мерной кубической решетки: деление по каждой переменной на равные промежутки.
Однако n - мерный куб имеет 2^n вершин, в отличие от n -мерного симплекса, имеющего $n+1$ вершину, что при больших n приводит к большому объему вычислений.
Напимер, при минимизации функции 10-и параметров нужно вычислить 11 значений для исходного симплекса и $2^{10} = 1024$ значения для исходного куба. При инвертировании симплекса нужно вычислить одно дополнительное значение, а у куба 2^{n-1} новых значений. (При $n = 10$ их 512).

Метод покоординатного спуска

Из некоторой начальной точки с координатами $x_1^0, x_2^0, \ldots, x_n^0$ ищется минимум функции одной переменной x_1 при фиксированных значениях $x_2^0, \ldots, x_n^0$, затем уже при новом значении x^1 и фиксированных $x_3^0, \ldots, x_n^0$ ищется одномерный минимум по x^2 и т.д.
Метод универсален, однако учет специфики задачи может позволить менять не одну, а сразу несколько координат для ускорения процесса минимизации.

Градиентный метод

Осуществляется одномерная минимизация по переменной t при изменении координат $\Delta x^i = t\, n^i$, где вектор n^i определяется "разностным градиентом":

$$n^i = -a^i\, \frac{f(x^1, \ldots, x^i + h^i, \ldots, x^n) - f(x^1, \ldots, x^n)}{h^i}; \quad i = 1, \ldots, n.$$

Константы h^i, a^i определяются из физических соображений, согласуясь с размерностью переменных x^i. Можно совершить несколько шагов с одним и тем же значением n^i для поиска минимума по t. После этого вычисляются новые значения вектора n^i.

Этот метод более прозрачен, когда переменные x^i приведены к безразмерному виду и масштаб их выбран так, что влияние их изменений на изменение минимизируемой функции приблизительно одинаково.

Метод наискорейшего спуска

Этот метод аналогичен предыдущему, но значения вектора n^i вычисляются заново после каждого шага по t.

Метод Хука-Дживса

Метод разработан в 1961 году. Поиск состоит из последовательности шагов вокруг базисной точки и затем – поиск по образцу.

- Выбор начальной базисной точки $\mathbf{r}_0 = (x_0^1, x_0^2, \ldots, x_0^n)$ и шага по каждой переменной h^i.
- Вычисляются значение функции и после приращения по каждой переменной по очереди на шаг h^i:

$$f_0 = f(\mathbf{r}_0); \quad f_i = f(x_0^1, x_0^2, \ldots, x_0^i + h^i, \ldots, x_0^n); \quad i = 1, \ldots, n.$$

- Если сдвиг приводит к увеличению значения, то шаг меняется на обратный:

$$if(f_i > f_0)h^i = -h^i, f_i = f(x_0^1, x_0^2, \ldots, x_0^i + h^i, \ldots, x_0^n).$$

- Если и после этого $f_i > f_0$, то полагается $h^i = 0$.

Таким образом формируется вектор сдвига $\mathbf{h} = (h^1, h^2, \ldots, h^n)$.

После этого начинается процедура *сдвига по образцу*

- Вычисляется значение функции $f(\mathbf{r}_1)$ в точке $\mathbf{r}_1 = \mathbf{r}_0 + \mathbf{h}$.
- Если $f(\mathbf{r}_1) < f(\mathbf{r}_0)$, то формируется новый вектор координат
$$\mathbf{r}_2 = \mathbf{r}_1 + \mathbf{h}. \tag{1}$$
  Такие шаги повторяются, пока функция не перестанет уменьшаться. Тогда предпоследняя точка вновь принимается за базисную, исходные шаги уменьшаются в два раза и весь процесс повторяется.
- Если шаги стали меньшими, чем установленная точность, процесс прекращается.

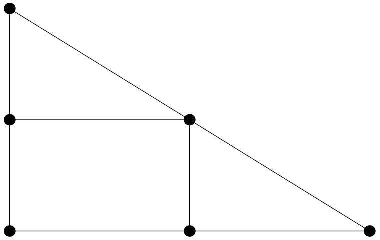
Метод квадратичной аппроксимации

Выбирается $(n+1)(n+2)/2$ точек и через них проводится квадратичная поверхность, у которой и ищется минимум.
Если в одномерном случае три точки для квадратичной аппроксимации можно выбирать достаточно произвольно, то в многомерном случае лучше этот выбор упорядочить.

Двумерная задача

Начнем с двумерной задачи, где нужно выбрать шесть точек для построения квадратичной аппроксимации. Если координаты начальной точки выберем за $(0,\, 0)$, остальные пять точек выбираем сдвигом от нее на один или два постоянных шага $h_x,\ h_y$:

$$
\begin{array}{lll}
r_{02} = (0,\, 2\,h_y); & & \\
r_{01} = (0,\, h_y); & r_{11} = (h_x,\, h_y); & \\
r_{00} = (0,\, 0); & r_{10} = (h_x,\, 0); & r_{20} = (2\,h_x,\, 0).
\end{array}
\tag{2}
$$



В этих точках находим значения минимизируемой функции $f_{00}, \ldots, f_{02}$. Теперь строим квадратичную двумерную аппроксимацию

$$F(r) = a_{11}x^2 + 2\, a_{12}\, x\, y + a_{22}\, y^2 - 2\, a_1\, x - 2\, a_2\, y + a_0 \tag{3}$$

и шесть коэффициентов a_0, a_i, a_{ij} находятся из шести линейных уравнений $F(r_i) = f_i$:

$$a_0 = f_{00}; \quad a_1 = \frac{3\, f_{00} - 4\, f_{10} + f_{20}}{4\, h_1}; \quad a_2 = \frac{3\, f_{00} - 4\, f_{01} + f_{02}}{4\, h_2};$$

$$a_{11} = \frac{f_{00} - 2f_{10} + f_{20}}{2\, h_1^2}; \quad a_{12} = \frac{f_{00} - f_{10} - f_{01} + f_{11}}{2\, h_1\, h_2}; \tag{4}$$

$$a_{22} = \frac{f_{00} - 2f_{01} + f_{02}}{2\, h_2^2}.$$

Теперь ищется минимум функции F:

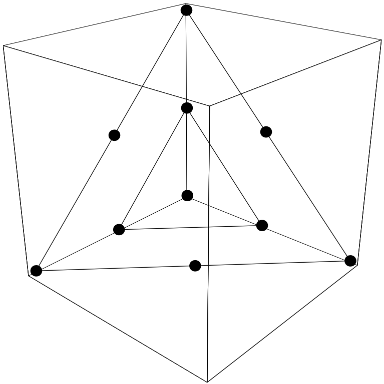
$$\frac{\partial F}{\partial x} = 2(a_{11}x + a_{12}y - a_1) = 0; \quad \frac{\partial F}{\partial y} = 2(a_{12}x + a_{22}\,y - a_2) = 0.$$

Из решения этой системы находится положение экстремума аппроксимирующей функции F (но не f).
Теперь нужно уменьшить шаги h_1 h_2 и повторить алгоритм в окрестности точки минимума x_0, y_0.

Трехмерная задача

В трехмерной задаче точки (2) получат еще третью координату $z = 0$ и к ним добавятся еще точки со смещением по z на один или два шага h_z:

$$r_{001} = (0,\ 0,\ h_z) \qquad r_{101} = (h_x,\ 0,\ h_z)$$
$$r_{011} = (0,\ h_y,\ h_z)$$
$$r_{002} = (0,\ 0,\ 2\,h_z)$$



По значениям минимизируемой функции в этих точках находятся коэффициенты квадратичной функции F. К коэффициентам (4) добавятся коэффициенты

$$a_3 = \frac{3\, f_{000} - 4\, f_{001} + f_{002}}{4\, h_3}; \quad a_{33} = \frac{f_{000} - 2 f_{001} + f_{002}}{2\, h_3^2};$$

$$a_{13} = \frac{f_{000} - f_{100} - f_{001} + f_{101}}{2\, h_1\, h_3}; \quad a_{23} = \frac{f_{000} - f_{010} - f_{001} + f_{011}}{2\, h_2\, h_3},$$

после чего сдвиг по x, y, z находится из системы линейных уравнений

$$a_{11}\Delta x + a_{12}\Delta y + a_{13}\Delta z = a_1;$$

$$a_{12}\Delta x + a_{22}\Delta y + a_{23}\Delta z = a_2;$$

$$a_{13}\Delta x + a_{23}\Delta y + a_{33}\Delta z = a_3.$$

Аналогично происходит добавление коэффициентов и уравнений при увеличении размерности.

Многомерная задача

В общем случае при наличии n переменных выполняются действия, которые можно было проследить на двумерной и трехмерной задаче.

Введем некоторые обозначения. Значение функции в исходной точке обозначается

$$f_0 = f(x^1, x^2, \ldots, x^n).$$

В точке, сдвинутой на один шаг по одной i-й переменной:

$$f_i = f(x^1, \ldots, x^i + h^i, \ldots, x^n).$$

В точке, сдвинутой на один шаг по двум разным переменным:

$$f_{ij} = f(x^1, \ldots, x^i + h^i, \ldots, x^j + h^j, \ldots, x^n); \quad i \leq j,$$

причем при $i = j$ функция вычисляется в точке, сдвинутой на два шага по одной переменной:

$$f_{ii} = f(x^1, \ldots, x^i + 2\,h^i, \ldots, x^n).$$

- Строится симметричная матрица a_{ij}:

$$a_{ij} = f_0 - f_i - f_j + f_{ij}, \quad i, j \le n. \tag{5}$$

- Находятся n коэффициентов b_i по формуле

$$b_i = \frac{3\, f_0 - 4\, f_i + f_{ii}}{2}. \tag{6}$$

- Решается система линейных уравнений

$$\sum_{j=1}^{n} a_{ij} \Delta^j = b^i. \tag{7}$$

- Точка экстремума квадратично аппроксимирующей функции (новая точка поиска) имеет координаты

$$\bar{x}^i = x^i + h^i\, \Delta^i.$$

  Нужно убедиться, что мы нашли минимум, а не максимум: $f(\bar{x}) < f_0$.

Если $h^i \Delta^i < \bar{\Delta}^i$ – заданных абсолютных погрешностей, или $\Delta^i < \delta^i$ – заданных относительных погрешностей, то поиск минимума прекращается, иначе процесс повторяется сначала, может быть, с измененными (уменьшенными) шагами.