Градиент и оптимизация гладких функций

1. Частные производные и градиент

Ранее было рассмотрено понятие «*производная*» для функции одной переменной, а также был изучен ее геометрический смысл. В случае функции двух и более переменных используются **частные производные**:

$$\frac{\Delta f}{\Delta x} \to f'_x, \quad x \to 0, \quad y \text{ — фиксирован;}$$

$$\frac{\Delta f}{\Delta y} \to f'_y, \quad y \to 0, \quad x \text{ — фиксирован.}$$

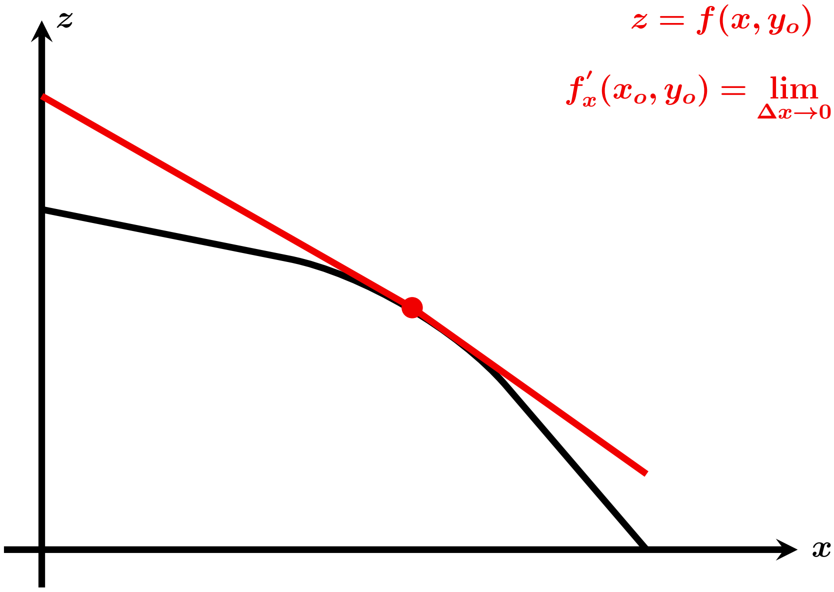

Рис. 1.

Геометрический смысл частных производных заключается в том, что график функции двух переменных приближается *касательной плоскостью*. В проекции на плоскость zx видно, что производная по x — это угловой коэффициент касательной (см. рис. 1). Аналогично в проекции на плоскость xy (см. рис. 2).

В пространстве график функции $f(x, y)$ приближается касательной плоскостью. Можно считать, что при переходе от точки (x_0, y_0) к точке (x, y) сначала координата x поменяется на Δx, а затем — y на Δy. Тогда изменение координаты z вдоль касательной плоскости запишется в виде:

$$\Delta f = f'_x \Delta x + f'_y \Delta y.$$

Вектор из частных производных функции называется **градиентом**:

$$\nabla f(x_0, y_0) = \begin{pmatrix} f'_x(x_0, y_0) \\ f'_y(x_0, y_0) \end{pmatrix}.$$

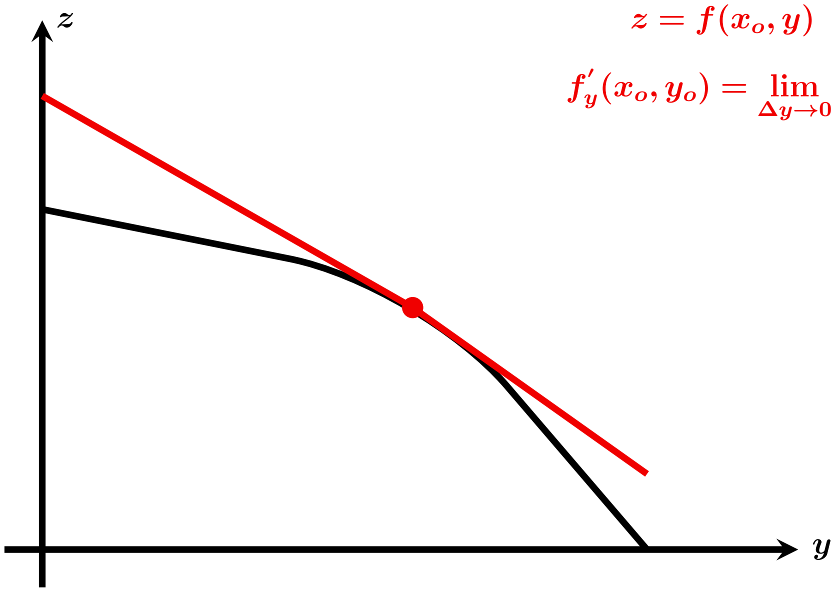

Рис. 2.

Градиент задает направление наискорейшего роста функции, а *антиградиент* (градиент, умноженный на -1) — направление наискорейшего убывания.

Другим способом визуализировать функцию двух переменных являются *линии уровня* — линии, вдоль которых функция принимает фиксированное значение. Изображая эти линии на плоскости, можно получить представление о том, как меняется функция в разных точках (см. рис. 3).

Рис. 3.

На рис. 4 изображен градиент вместе с линиями уровня.

В многомерном случае рассматривается функция n переменных:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \in \mathbb{R}^n, \quad f\big(x_1, \ldots, x_N\big) = f(x) \text{ — функция от вектора.}$$

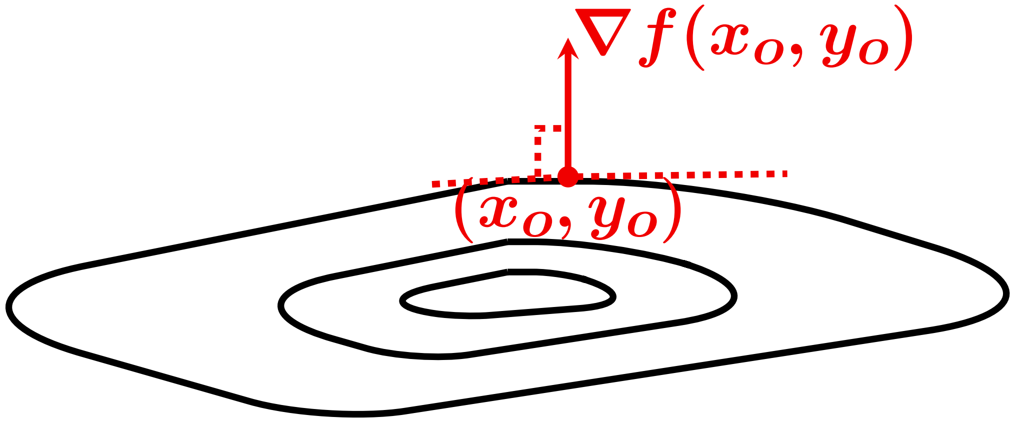

Рис. 4.

В случае, когда переменных две, можно использовать это же обозначение:

$$f(x, y) = f(x_1, x_2).$$

Является ли x вектором или скаляром, как правило, становится понятно из контекста.

2. Применение градиента

Пусть необходимо решить некоторую задачу оптимизации:

$$f\big(x_1, \dots, x_N\big) \to \min.$$

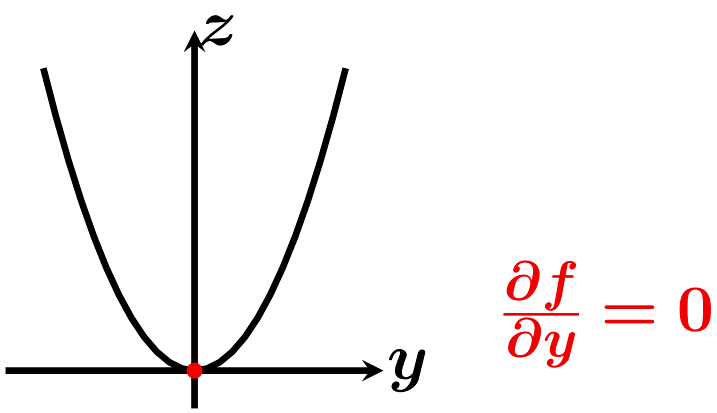

Рис. 5.

Например, такая задача возникает при выборе оптимальных параметров рекламной кампании или при настраивании параметров сложного технологического процесса (минимизируется вероятность неудачи). Возникает вопрос — как именно нужно выбрать параметры, чтобы функция была минимальной?

Если зафиксировать значение x или значение y в двумерном случае и рассмотреть функцию в координатах zy или zx соответственно, то станет понятно, что значения частных производных в точке экстремума должны быть нулевыми (см. рис. 5 и 6).

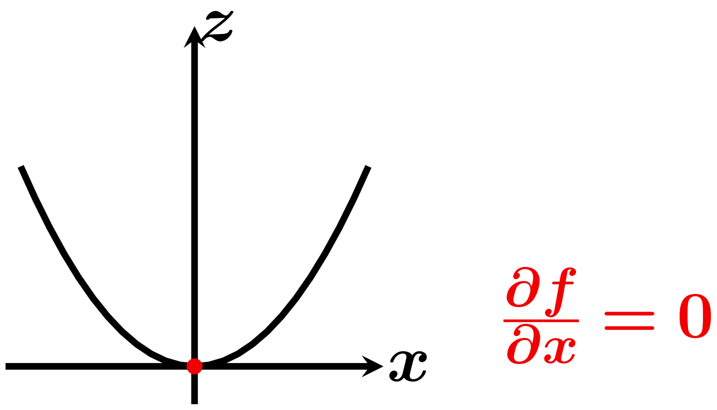

Рис. 6.

В силу равенства нулю всех частных производных:

$$\nabla f = 0.$$

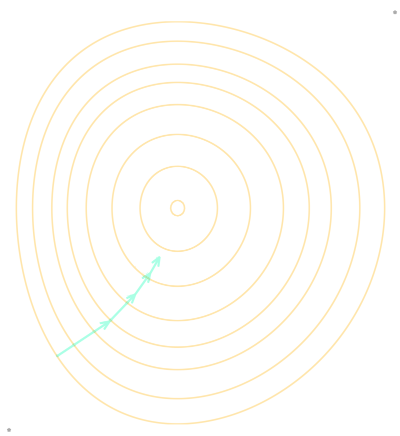

Рис. 7.

Если найти экстремум не получается аналитически, необходимо использовать другой подход — *численную оптимизацию*. Если известен градиент функции, то нужно начать в произвольной точке $x_1, \ldots, x_N$ и на каждом шаге немного смещаться в направлении антиградиента, постепенно приближаясь к минимуму (*градиентный спуск*, см. рис. 7):

$$x_{n+1} = x_n - \gamma_n \nabla f(x_n).$$

У данного метода имеется следующая аналогия. Человек приехал в незнакомое место и поселился в низине. Гуляя по недалеким окрестностям, он заблудился. Как ему скорее попасть домой? Очевидно, ему нужно двигаться по самому крутому уклону вниз.

Здесь возникает вопрос, насколько такой спуск будет безопасным? Если нужно найти экстремум при дополнительных условиях, то это задача *условного экстремума*.

3. Производная по направлению

Для того, чтобы узнать, как меняется функция в определенной точке x в направлении l, необходимо воспользоваться понятием **производная по направлению**:

$$\frac{\partial f}{\partial l} = \lim_{t \to 0} \frac{f(x + t \cdot l) - f(x)}{t},$$

где x и l — векторы, t — число.

Если интересует только направление вектора l, то можно ограничиться векторами длины 1 (см. рис. 8).

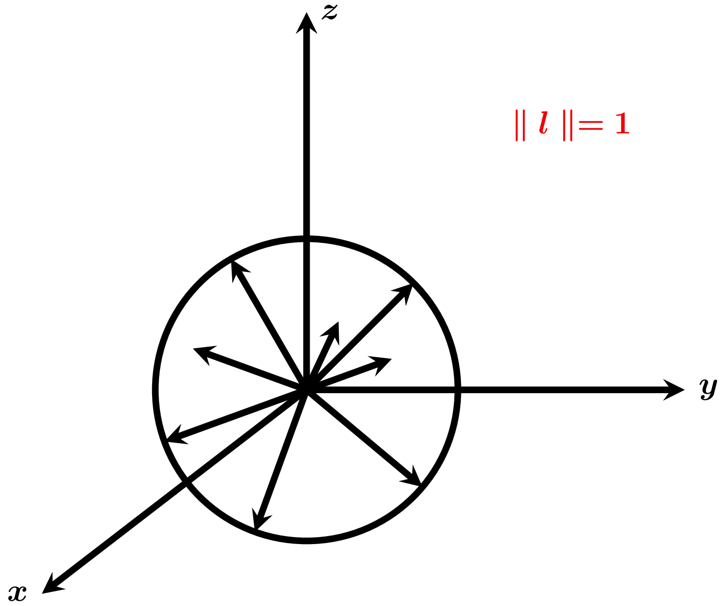

Рис. 8.

Производная по направлению имеет наглядный геометрический смысл, который проще всего пояснить в двумерном случае. Поверхность, являющаяся графиком функции $f(x, y)$, рассекается плоскостью, проходящей через точку (x_0, y_0) вдоль вектора l и перпендикулярно плоскости xy. В этом сечении можно рассматривать исходную функцию как функцию от параметра t (см. рис. 9). Тогда производная по направлению — это производная новой функции от параметра t в точке $t = 0$, а ее геометрический смысл — это скорость роста функции в заданном вектором l направлении.

4. Касательная плоскость и линейное приближение

Рассматривается изменение значения функции $f(x, y)$ при переходе от точки (x_0, y_0) к точке $(x_0 + \Delta x,\, y_0 + \Delta y)$:

$$\Delta f = f\big(x_0 + \Delta x,\, y_0 + \Delta y\big) - f\big(x_0, y_0\big).$$

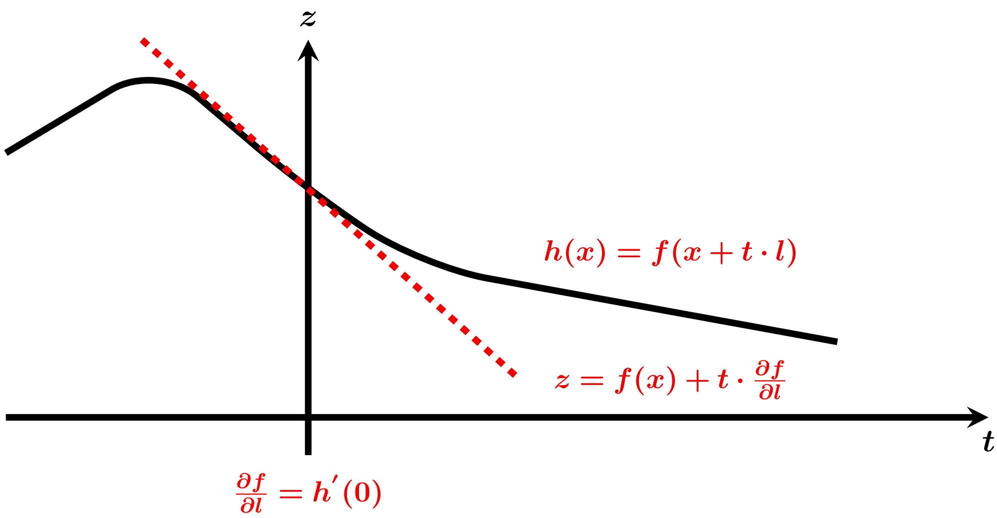

Рис. 9.

Приближенная формула для прироста имеет вид:

$$\Delta f \approx f'_x(x_0, y_0) \cdot \Delta x + f'_y(x_0, y_0) \cdot \Delta y.$$

Данную формулу можно записать компактнее, используя скалярное произведение:

$$\Delta f \approx \left\langle \nabla f(x_0, y_0), \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \right\rangle.$$

Эта формула описывает приближение графика функции касательной плоскостью (см. рис. 10). Уравнение касательной плоскости совпадает с приближенной формулой для прироста:

$$z = g(x, y) = \left\langle \nabla f(x_0, y_0), \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \right\rangle + f(x_0, y_0).$$

Рис. 10.

Необходимо заметить, что чем ближе к точке (x_0, y_0), тем точнее выполнено приближенное равенство. Так как данное выражение для Δf линейно по $\Delta x = x - x_0$

и $\Delta y = y - y_0$, оценка значений функции с помощью этого выражения называется *линейным приближением*:

$$f(x,y) \approx f\big(x_0,y_0\big) + \left\langle \nabla f\big(x_0,y_0\big), \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \right\rangle.$$

Линейное приближение часто используется для анализа сложных функций, в частности, для построения алгоритмов анализа данных.

5. Направление наискорейшего роста

Необходимо выяснить, как связана с градиентом производная по направлению. Будет рассматриваться двумерный случай:

$$l = \begin{pmatrix} l_x \\ l_y \end{pmatrix}, \quad \frac{\partial f}{\partial l}\big(x_0,y_0\big) = \lim_{t\to 0} \frac{f(x_0 + t\cdot l_x,\, y_0 + t\cdot l_y) - f(x_0,y_0)}{t}.$$

Вводятся следующие обозначения:

$$\Delta f = f\big(x_0 + t\cdot l_x,\, y_0 + t\cdot l_y\big) - f\big(x_0,y_0\big), \quad \Delta x = t\cdot l_x, \quad \Delta y = t\cdot l_y.$$

Используя приближенную формулу, можно записать:

$$\Delta f \approx \left\langle \nabla f\big(x_0,y_0\big), \begin{pmatrix} t\cdot l_x \\ t\cdot l_y \end{pmatrix} \right\rangle = t\cdot \Big\langle \nabla f\big(x_0,y_0\big), l \Big\rangle.$$

При стремлении t к нулю все ближе и ближе точка (x_0,y_0), поэтому можно воспользоваться приближенной формулой под знаком предела, что приведет к производной по направлению в этой точке. Далее нужно найти, в каком случае производная по направлению будет максимальной:

$$\frac{\partial p}{\partial l} = \Big\langle f\big(x_0,y_0\big), l \Big\rangle, \quad \langle f, l\rangle = ||\nabla f|| \cdot ||l|| \cdot \cos\theta \to \max.$$

Для этого угол θ между градиентом и вектором l должен быть равен нулю. Таким образом, вектор l должен быть сонаправлен с градиентом, следовательно, градиент задает направление наискорейшео роста.