Методы оптимизации

конспект лекций

Бойцев А.А., Попов А.М.
Университет ИТМО – Санкт-Петербург, 2025 г.

Содержание

Введение

Математическая оптимизация — это мощный инструмент, который лежит в основе решения множества задач, с которыми мы сталкиваемся как в науке, так и в повседневной жизни. Её суть заключается в поиске наилучшего решения из множества возможных, будь то минимизация затрат, максимизация прибыли, повышение эффективности или достижение оптимального баланса между различными параметрами. В современном мире, где данные и ресурсы часто ограничены, умение находить оптимальные решения становится критически важным навыком.

Математическая оптимизация применяется в самых разных областях: от экономики и инженерии до машинного обучения.

Но зачем изучать оптимизацию, если многие современные инструменты и библиотеки уже предоставляют готовые решения? Ответ прост: понимание принципов оптимизации позволяет не только использовать существующие методы, но и разрабатывать новые, адаптированные под конкретные задачи. Это особенно важно в ситуациях, когда стандартные подходы оказываются неэффективными или неприменимыми.

Данный раздел математики очень богат, и включает дискретную, непрерывную, смешанную оптимизации, глобальную и локальную, безусловную и условную, и многое другое. В рамках курса мы затронем только непрерывный случай, локальный и без условий.

1 Терминология

1.1 Задача оптимизации

Будем рассматривать следующую постановку задачи:

Есть некоторая функция $f : X \to \mathbb{R}$. Искать будем значения $x^* \in X$ такие, что

$$f(x^*) = \min_{x \in X} f(x).$$

Допущение. *В рамках курса, за X будет скрываться некоторое подмножество $\mathbb{R}^n$, а функция f будет на нём непрерывной.*

Напомним некоторые определения, известные читателю из курса математического анализа.

Определение 1. *Пусть $f : X \to \mathbb{R}$, $x_0 \in X$. Точка x_0 называется точкой локального минимума функции f, если*

$$\exists \varepsilon > 0 : \quad \forall x \in X : \ ||x - x_0|| < \varepsilon \quad \Rightarrow \quad f(x) \geqslant f(x_0)$$

Определение 2. *Пусть $f : X \to \mathbb{R}$, $x_0 \in X$. Точка x_0 называется точкой глобального минимума функции f, если*

$$\forall x \in X \quad \Rightarrow \quad f(x) \geqslant f(x_0)$$

Верные остается, что точка глобального минимума является точкой локального. Обратное верно не всегда.

Замечание. *Методы, которые мы будем затрагивать, буду искать локальные минимумы. Поэтому, при решении задачи*

$$f(x) \longrightarrow \min, \quad x \in X$$

мы не будем принебрегать локальными минимумами.

Замечание. *Аналогично тому, мы можем рассматривать задачу максимизации. Поиск точек максимума для f полностью переносится на поиск точек минимума для $-f$.*

1.2 Задача безусловной оптимизации

В случае, когда $X = \mathbb{R}^n$, поиск точки минимума функции f становится задачей безусловной оптимизации.

Введем пару обозначений:

- Градиент или же просто производная

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

- Гессиан или матрица вторых производных

$$\nabla^2 f(x) = \left\{\frac{\partial^2 f}{\partial x_i \partial x_j}\right\}_{i,j=1}^{n} = f''(x)$$

Напомним, что функции, обладающие достаточной гладкостью, могут быть приближены формулой Тейлора

$$f(x_0 + h) = f(x_0) + \nabla f(x)^T h + h^T \nabla^2 f(x_0) h + o(||h||^2),$$

где $o(||h||^2) = \alpha(h)||h||^2$, $\alpha(h) \to 0$ при $h \to 0$.

Напомним две важные теоремы из курса анализа:

Теорема 1 (Необходимое условия минимума)**.** *Пусть $f : \mathbb{R}^n \to \mathbb{R}$ – дифференцируемая функция. Если x_0 – точка минимума, то*

$$\nabla f(x_0) = f'(x_0) = 0.$$

Доказательство. Пусть x_0 – точка локального минимума. Тогда для любого $h \in \mathbb{R}^n$ и достаточно малого $\varepsilon > 0$ выполнено

$$0 \leqslant f(x_0 + \varepsilon h) - f(x_0) = f'(x_0)^T \varepsilon h + o(\varepsilon).$$

Разделим на ε и устремим его к нулю, тогда получим

$$f'(x_0)^T h \geqslant 0.$$

Требуемое получим, если положим $h = -f'(x_0)$.

□

Замечание. *Соответствующая теорема справедлива и в случае, когда $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^n$, и x_0 – внутренняя для X.*

Теорема 2 (Достаточное условие экстремума)**.** *Пусть $f \in C^2(\mathbb{R}^n)$ и x_0 – точка экстремума. Если $f''(x_0)$ положительно определена, то x_0 – точка локального минимума.*

Замечание. *Положительная определенность означает, что*

$$h^T f''(x_0) h > 0 \quad \textit{при} \quad h \in \mathbb{R}^n \setminus \{0\}.$$

Доказательство. Предположим противное, то есть существует x_k такая, что $x_k \to x_0$, $x_k \neq x_0$ и $f(x_k) \leqslant f(x_0)$. Представим

$$x_k = x_0 + \alpha_k h_k, \quad \alpha_k = \|x_k - x_0\|, \quad h_k = \frac{x_k - x_0}{\alpha_k}.$$

Ясно, что $h_k \to h \neq 0$. Так как $f'(x_0) = 0$, то

$$0 \geqslant f(x_k) - f(x_0) = \frac{1}{2}\alpha_k^2 h_k^T f''(x_0) h_k + o(\alpha_k^2).$$

Разделив на α_k^2 и перейдя к пределу, придем к противоречию. □

1.3 Оракул

Определение 3. *Оракул – блок кода, отвечающий за вычисление информации о задаче.*

Типичный пример оракула нулевого порядка – это $O(x) = \{f(x)\}$. С ним мы можем вычислять только значение функции в произвольной точке. $O(x) = \{f(x), f'(x)\}$ – оракул первого порядка, позволяющий узнать как значение функции, так и значение производной этой функции. Оракул второго порядка – это $O(x) = \{f(x), f'(x), f''(x)\}$. Можно продолжать. Оракулы порядка выше второго редко используются на практике.

Оракулы могут быть и более хитрой формы, например, стохастический. Пусть

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x).$$

Понятно, что если n велико, то вычисление оракула даже первого порядка очень затруднительно. Стохастический оракул: $O(x) = \{f_{i_k}(x), f'_{i_k}(x)\}$, $i_k \in \{1, 2, ..., n\}$ – случайно.

1.4 Общая схема итерационного алгоритма оптимизации

Общую схему можно изобразить так:

1. Задаемся начальным приближением x_0 и погрешностью ε.
2. Пусть $I_{-1} = \varnothing$ – начальная информация.
3. Для $k \in \{0, 1, 2, ...\}$
   (a) Вызываем $O(x_k)$
   (b) $I_k = I_{k-1} \cup \{x_k, O(x_k)\}$ – накапливаем информацию

(c) Вычисляем x_{k+1} по информации I_k.

(d) Проверяем критерий останова (зависит от ε).

Критерии могут быть разные, например: $|f(x_{k+1})-f(x_k)| < \varepsilon$, $\|x_{k+1}-x_k\| < \varepsilon$, $\|f'(x_{k+1})\| < \varepsilon$ и так далее. Обязательно должно быть ограничено количество итераций.

Замечание. *При реализации оракула полезно себя проверять! Например, для проверки реализации оракула первого порядка, можно использовать разностное приближение производной:*

$$\frac{\partial f}{\partial x_i}(x) \approx \frac{f(x+\varepsilon e_i) - f(x)}{\varepsilon}.$$

Слишком маленькое ε брать нельзя – проблема с точностью. $\varepsilon \approx \sqrt{\varepsilon_m}$ – корень из машинной точности. В числах типа double $\varepsilon_m \approx 10^{-16}$.

Скажем пару слов о критерии останова. Типично, момент остановки определяется следующими условиями:

1. Количество итераций превосходит заданный N;
2. $||x_{n+1} - x_n|| < \varepsilon$ - шаг точки становится слишком малым;
3. $||f(x_{n+1}) - f(x_n)|| < \varepsilon$ - функция слишком мало меняется;
4. $||\nabla f(x_n)|| < \varepsilon$ - градиент слишком мал (а значит, рядом его ноль).

Этот список не является полным. Например, можно составлять условия в виде линейной комбинации левых частей условий 2-4.

1.5 Скорость сходимости итерационных процессов

Определение 4. *Обозначим за $r_k = ||x_k - x^*||$ - норму разности между значением на текущей итерации и решением задачи. Метод называется сходящимся, если $x_k \to x^*$ при $k \to \infty$, или, равносильно $r_k \to 0$.*

Возникает вопрос, а как оценить скорость сходимости, ведь стремление r_k к нулю может быть разным. Введем дополнительные определения, чтобы дать ответ на вопрос.

Определение 5. *Говорят, что r_k сходится линейно, если*

$$\exists C \in (0,1): \ \forall k > k_0 \ \ r_{k+1} \leqslant C r_k.$$

Это равносильно тому, что

$$\lim \sup_{k\to+\infty} \frac{r_{k+1}}{r_k} = C \in (0,1).$$

Но почему такая скорость сходимости называется линейной? Это становится очевидным, после рассмотрения логарифма от r_k

$$r_k \leqslant C^k r_0.$$

Тогда

$$\log r_k \leqslant k \log c + \log r_0.$$

Определение 6. *Говорят, что r_k сходится сублинейно, если*

$$\forall k > k_0 \ r_{k+1} \leqslant r_k.$$

Это равносильно тому, что

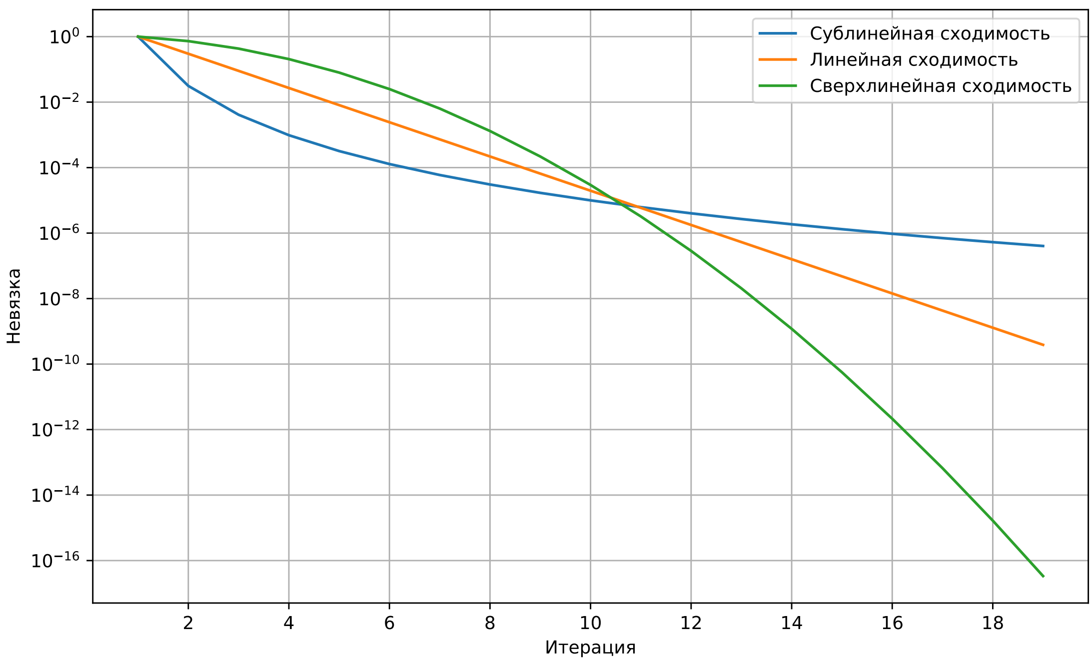
$$\lim_{k \to +\infty} \sup \frac{r_{k+1}}{r_k} = 1.$$

Определение 7. *Говорят, что r_k сходится сверхлинейно, если*

$$\forall k > k_0 \ r_{k+1} \leqslant C_k r_k, \ \text{где } C_k \xrightarrow[k \to 0]{} 0.$$

Это равносильно тому, что

$$\lim_{k \to +\infty} \sup \frac{r_{k+1}}{r_k} = 0.$$



Сравнение скоростей сходимости

Определение 8. *Говорят, что сходимость r_k квадратичная, если*

$$\exists C \in (0, 1) \ \forall k > k_0 \ r_{k+1} \leqslant C r_k^2$$

Замечание. *Если сходящаяся последовательность r_k не является монотонно убывающей, т.е. для любого наперед заданного $\forall k_0 \ \exists k > k_0 \ : \ r_{k+1} > r_k$, для утверждения о скорости достаточно оценить её сверху монотонной.*

2 Оптимизация нулевого порядка

Начнём в каком-то смысле с наиболее простого случая – нам необходимо минимизировать функцию одного аргумента, не используя производные. Это часто встречающаяся задача редко появляется обособлено, а о её приложениях вы узнаете в следующем разделе.

2.1 Унимодальные функции

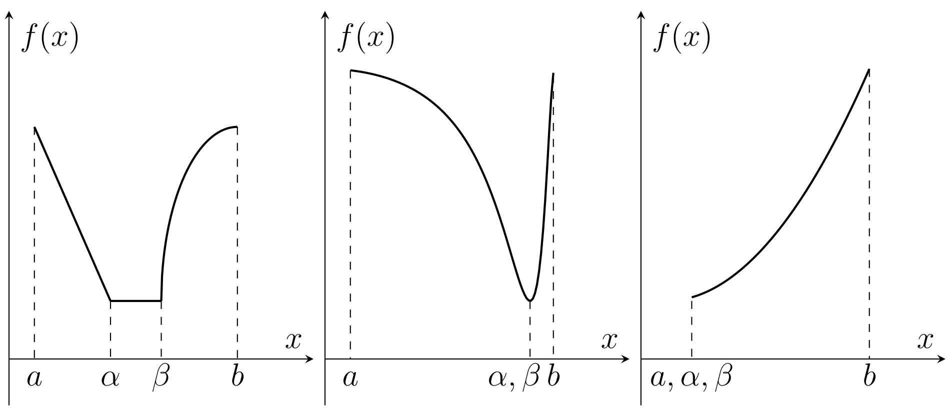
Будем решать задачу следующего вида

$$f(x) \to \min, \quad x \in [a, b],$$

, причем будем предполагать функцию f непрерывной и унимодальной.

Определение 9. *Функция $f(x) : [a, b] \to \mathbb{R}$ называется унимодальной, если существуют такие α, β, причем $a \leqslant \alpha \leqslant \beta \leqslant b$, что:*

1. *на $[a, \alpha]$ функция f монотонно убывает;*
2. *на $[\beta, b]$ функция f монотонно возрастает;*
3. *на $[\alpha, \beta]$ функция постоянна, более того $f(y) = \min\limits_{x \in [a,b]} f(x)$ при $y \in [\alpha, \beta]$*



Примеры графика унимодальной функции

Замечание. *Часто, под унимодальной функцией подразумевают те функции, которые обладают единственной точкой минимума, что в нашем определении равносильно $\alpha = \beta$.*

Можно заметить, что на первом из предложенных рисунков точка минимума не единственная. Для некоторых методов это не станет препятствием, и мы просто найдем одну из возможных. Для других методов это станет настоящей проблемой.

Мы попробуем итеративным путём уменьшать размер отрезка, до тех пор, пока нас не устроит его длина, или до того момента, как значения функций будет меньться слишком слабо. Выберем две точки x и y, такие, что $a < x < y < b$. Сформулируем общий алгоритм для шага итерации:

1. Если $f(x) > f(y)$, то минимум находится на отрезке $[x, b]$.
2. Если $f(x) < f(y)$, то минимум находится на отрезке $[a, y]$.
3. Если $f(x) = f(y)$, то минимум находится на $[x, y]$ (редко на практике).

Логичным станет вопрос о выборе x и y.

2.2 Метод золотого сечения

Предложим метод, основанный на следующих вполне разумных допущениях:

1. Количество вычислений функции минимально;
2. Рассматриваемые отрезки $[a, y]$ и $[x, b]$ имеют одинаковую длинну;
3. Сокращение отрезка происходит в одну и ту же величину.

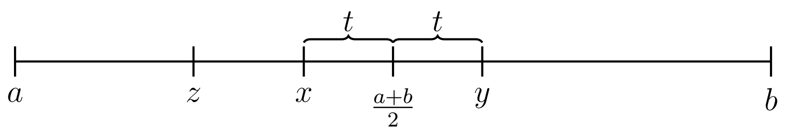


Рис. 1: Точки первых двух итераций

Рассмотрим две итерации и выясним, на какую величину сокращается длПусть на первой итерации выбран отрезок $[a, y]$. Тогда, ради выполнения первого условия, одной из рассматриваемых точек второй итерации должен стать x (ведь значение функции мы уже вычислили, когда сравнивали $f(x)$ и $f(y)$). В силу второго условия, точки x и y находятся симметрично относительно центра отрезка. А третье условие дает следующую пропорцию:

$$\frac{l[a, y]}{l[a, b]} = \frac{l[a, x]}{l[a, y]}.$$

Считая, что длинна $l[a, b] = 1$, а расстояние от центра отрезка до x и до y равно t:

$$\frac{\frac{1}{2} + t}{1} = \frac{\frac{1}{2} - t}{\frac{1}{2} + t}.$$

Решая уравнение, с учётом того, что ответ положительный:

$$t = \frac{-2 + \sqrt{5}}{2}.$$

Таким образом, получаем коэффициент пропорции:

$$K = \frac{l[a, y]}{l[a, b]} = 1 + t = \frac{\sqrt{5} - 1}{2}.$$

Замечание. *Метод с таким коэффициентом K называется методом золотого сечения не просто так. Ведь $\Phi = \frac{1}{K} = \frac{1+\sqrt{5}}{2}$ и есть золотая пропорция!*

Сформулируем алгоритм:

Алгоритм 1 Алгоритм золотого сечения

Вход: Функция f, отрезок $[a, b]$, точность ε, максимальное количество итераций N.
Выход: Точка минимума $x_{\min}$ и значение в ней $f(x_{\min})$.

Инициализация:

$$K := \frac{\sqrt{5} - 1}{2}, \quad L_0 := b - a, \quad L_1 := KL_0, \quad x := b - L_1, \quad y := a + L_1.$$

Вычислить $f_x = f(x)$, $f_y = f(y)$.
Для $k \in \{1, 2, ..., N\}$ **Выполнять**
    $L_{k+1} = KL_k$.
    Если $f_x < f_y$ **тогда**
        $b := y, \quad y := x, \quad f_y := f_x, \quad x = b - L_{k+1}$.
        Вычислить $f_x := f(x)$.
    иначе
        $a := x, \quad x := y, \quad f_x := f_y, \quad y := a + L_{k+1}$
        Вычислить $f_y := f(y)$.
    Конец условия
    Если $L_{k+1} < \varepsilon$ **тогда**
        $x_{\min} := \arg\min(f_x, f_y)$, $f(x_{\min}) := \min(f_x, f_y)$
        Выход из цикла.
    Конец условия
Конец цикла

Оценим скорость алгоритма. Очевидно, что она линейная, в силу неравенства

$$L_N = K^N L_0 < \varepsilon,$$

переписываемого с логарифмом в виде

$$N \log K + \log L_0 < \log \varepsilon \;\Rightarrow\; N > \frac{\log \varepsilon}{\log K} - \frac{\log L_0}{\log K}, \quad \log K \approx -0.2.$$

Видно, что коэффициент при $\log \varepsilon$ равен примерно -5, что хорошо: требуется около пяти итераций на улучшение одной значащей цифры. Например, если $L_0 = 10^2$, $\varepsilon = 10^{-4}$, то нужно около 30 итераций.

Замечание. *Если функция не унимодальна, то метод может сойтись к локальному минимуму, а может не сойтись вовсе. В общем случае можно говорить о такой эвристике: если посчитаны значения функции в трех точках, и в «средней» точке значение меньше, чем в крайних, то метод сойдется. Иначе – не факт.*

2.3 Метод парабол

Будем приближать функцию моделью – параболой. Пусть есть три точки x_1, x_2, x_3. Построим параболу через три точки, ищем ее минимум и получаем четвертую точку. Дальше сужаем промежуток и повторяем все сказанное.

Пусть $g(x) = \alpha x^2 + \beta x + \gamma$, тогда

$$f(x_1) = g(x_1), \quad f(x_2) = g(x_2), \quad f(x_3) = g(x_3).$$

На практике систему лучше решать программно, полезно вызвать линейный солвер (для python, это, например numpy.linalg.solve).

$$x_4 = \underset{x \in [\min x_i,\ \max x_i]}{\arg\min}\, g(x).$$

Конечно, нам нужно, чтобы минимум попал в наш изначальный отрезок. Для этого можно пользоваться той же эвристикой, что была предложена в методе золотого сечения. К сожалению, на поиск трех точек, удовлетворяющих условиям, надо потратить время и вычислительные силы.

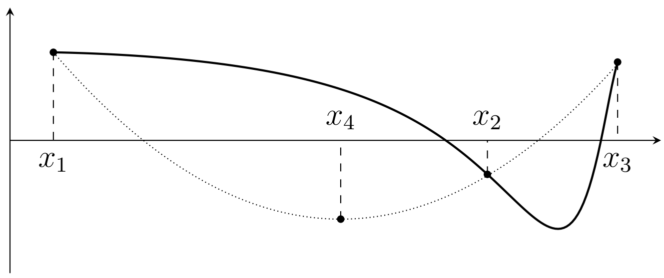


Рис. 2: Поиск x_4 по x_1, x_2, x_3

Алгоритм 2 Алгоритм метода парабол

Вход: Функция f, отрезок $[a, b]$, точность ε, максимальное количество итераций N, $a < c < b$, $f(a) > f(c)$, $f(b) > f(c)$.

Выход: Точка минимума $x_{\min}$ и значение в ней $f(x_{\min})$.

Инициализация:

$$f_a := f(a), \quad f_b := f(b), \quad f_c := f(c).$$

Для $k \in \{1, 2, ..., N\}$ **Выполнять**
- Найти точку u – вершину методом парабол.
- Вычислить $f_u = f(u)$
- **Если** $f_u < f_c$, $u < c$ **тогда**
  - $b := c,\ f_b := f_c,\ c := u,\ f_c := f_u.$
- **иначе если** $f_u > f_c$, $u < c$
  - $a := u,\ f_a := f_u.$
- **иначе если** $f_u < f_c$, $u > c$
  - $a := c,\ f_a := f_c,\ c := u,\ f_c := f_u.$
- **иначе**
  - $b := u,\ f_b := u$
- **Конец условия**
- **Если** $|a - c| < \varepsilon$, **тогда**
  - $x_{min} := \arg\min(f_a, f_b),\ f(x_{min}) := \min(f_a, f_b).$
  - Выход из цикла.
- **Конец условия**

Конец цикла

Метод парабол локально обладает суперлинейной скоростью сходимости. Это значит, что если стартовать рядом с точкой оптимума, то скорость будет суперлинейна. Если где-то в другом месте – нет никаких гарантий.

Замечание. *Для метода парабол очень важно, чтобы точки (a, f_a), (b, f_b) и (c, f_c) не оказались на одной прямой – наша модельная парабола станет вырожденной. Поэтому, для унимодальный функций, в которых $\alpha < \beta$ (есть некоторая планка с постоянным значением) метод не подойдёт.*

2.4 Метод Брента

На практике используется комбинированный метод (например, в пакете scipy.optimize). Идея следующая: пробуем метод парабол. Если промежуток начинает сжиматься реально суперлинейно, то и хорошо, продолжаем параболы. Иначе – действуем согласно золотому сечению. Сразу приведем алгоритм.

Алгоритм 3 Алгоритм метода Брента

Вход: Функция f и точки a, b, c, $f(c) < f(a)$, $f(c) < f(b)$. **Выход:** Точка минимума $x_{\min}$ и значение в ней $f(x_{\min})$.

1: Инициализация:

$$x := c, \quad w := c, \quad v := c, \quad f_x := f(c), \quad f_w := f(c), \quad f_v := f(c).$$

2: **Для** $k \in \{1, 2, ..., N\}$ **Выполнять**
3: **Если** значения f_x, f_w, f_v разные, **тогда**
4: Применить общий метод парабол, найти u.
5: **Конец условия**
6: **Если** $u = \mathsf{NaN}$, $u \notin [a, b]$ или $|u - x| > |v - b|/2$, **тогда**
7: Найти u методом золотого сечения большего из отрезков $[a, x]$ и $[x, b]$.
8: **Конец условия**
9: Вычислить $f_u = f(u)$.
10: Переопределить переменные
11: **Конец цикла**

Итак, алгоритм простым языком описать можно, например, так. Изначально на вход подается интервал, на котором расположен минимум – это интервал (a, b), а также точка c такая, что как алгоритм золотого сечения, так и метод парабол оказываются разумными для применения. x – текущее лучшее приближение, w – предыдущее лучшее приближение, v – предпредыдущее лучшее приближение; f_x, f_w, f_v – значения в лучших приближения, соответственно. Метод работает так: если есть возможность применить метод парабол, то метод применяется. Далее смотрим, удовлетворяет ли полученная точка условиям: лежит ли в отрезке, и достаточно ли быстро улучшается (приближается) новая точка относительно старой (половина длины предпредыдущего улучшения). Если все хорошо, действуем стандартно. Если нет, находим точку методом золотого сечения и действуем, опять же, стандартно.

3 Градиентный спуск

3.1 Основные понятия и схема метода

Метод градиентного спуска – достаточно простой метод, который можно получить из разных соображений. например, можно получить следующим образом.

Итак, локально рядом с точкой x_0 у достаточно гладкой функции справедливо следующее разложение:

$$f(x) \approx f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2!}(x - x_0)^T \nabla^2 f(x_0)(x - x_0) + ...$$

Написанное – не что иное, как формула Тейлора. Обрывая эту формулу на каком-то шаге, мы и получаем интересующую нас модель. Обрубим первыми членами, получим:

$$f(x) \approx f(x_0) + \nabla f(x)^T (x - x_0).$$

Искать минимум выражения справа плохо. Зададим условие, что $\|x - x_0\| \leqslant \varepsilon$, тогда задача

$$\nabla f(x)^T (x - x_0) \to \min$$

имеет очевидное решение:

$$x - x_0 = -\frac{\nabla f}{\|\nabla f\|}\varepsilon \ \Rightarrow \ x_{i+1} = x_i - \alpha \nabla f(x_i).$$

Это и есть схема градиентного спуска.

Алгоритм 4 Метод градиентного спуска

Вход: Функция f, начальное приближение x_0, точность ε, максимальное количество итераций N, стратегия выбора α.

```
Для k ∈ {1, 2, ..., N} Выполнять
    Вычислить f'(x_0) = ∇f(x_0).
    Выбрать α согласно стратегии.
    x := x_0 − αf'(x_0).
    Если Выполнен критерий остановки тогда
        Возвратить x, f(x).
    Конец условия
    x_0 := x
Конец цикла
```

Выход: Точка минимума $x_{\min}$ и значение в ней $f(x_{\min})$.

Остаётся незатронутым важный вопрос: что за стратегии выбора шага α. Существует несколько подходов – можно решать задачу одномерной оптимизации на каждом шаге - искать такое α, что $f(x_i - \alpha \nabla f(x_i))$ минимальна на каком-то промежутке, можно подобрать какую-то константу, а можно прибегнуть к так называемым условиям Армихо-Вульфа.

3.2 Липшицевы функции

Определение 10. *Говорят, что функция f принадлежит классу $C_L^{k,m}(X)$, если*

$$f \in C^k(X), \quad \|\nabla^m f(x) - \nabla^m f(y)\| \leqslant L\|x - y\|,$$

L при этом называется константной Липшица.

Нам, в большинстве своем, будут интересны так называемые функции с липшицевым градиентом, то есть функции класса $C_L^{1,1}(X)$. Отметим их дифференциальные свойства.

Теорема 3. *Пусть $f \in C_L^{1,1}(\mathbb{R}^n)$. Тогда $\forall x, y \in \mathbb{R}^n$ справедливы неравенства*

$$f(y) \leqslant f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2,$$

$$f(y) \geqslant f(x) + \nabla f(x)^T(y - x) - \frac{L}{2}\|y - x\|^2,$$

Доказательство.

$$f(y) = f(x) + \int\limits_0^1 f'(x + \tau(y - x))^T(y - x)d\tau =$$

$$= f(x) + f'(x)^T(y - x) + \int\limits_0^1 \left(f'(x + \tau(y - x)) - f'(x)\right)^T (y - x)d\tau.$$

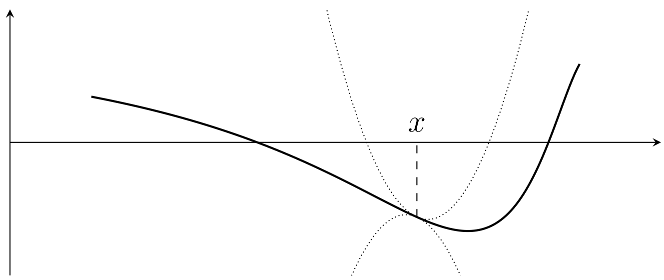
Тогда,

$$\left|f(y) - f(x) - f'(x)^T(y - x)\right| \leqslant \int\limits_0^1 \left|\left(f'(x + \tau(y - x)) - f'(x)\right)^T (y - x)\right| d\tau \leqslant$$

$$\int\limits_0^1 \|f'(x + \tau(y - x)) - f'(x)\|\|y - x\|d\tau \leqslant \int\limits_0^1 \tau L\|y - x\|^2 d\tau = \frac{L}{2}\|y - x\|^2.$$

□

Геометрически последнее условие означает, что в каждой точке функция может быть приближена сверху и снизу параболами, направленными в разные стороны, имеющими одинаковую кривизну.



Замечание. *Понятно, что в случае, когда $f \in C^2(\mathbb{R}^n)$, то $L = \sup\limits_{x \in \mathbb{R}^n} \alpha_{\max}(\nabla^2 f(x))$*

3.3 Градиентный спуск для $C_L^{1,1}$ функций

Пусть $f \in C_L^{1,1}$. Напомним, что справедливо следующее неравенство

$$f(y) \leqslant f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2.$$

Пусть y – шаг нашего метода, то есть $y = x - \alpha \nabla f(x)$. Тогда

$$f(y) \leqslant f(x) - \alpha\|\nabla f(x)\|^2 + \alpha^2 \frac{L}{2}\|\nabla f(x)\|^2.$$

Мы хотим как можно мощнее уменьшить f. Тогда можно подобрать такое α, чтобы правая часть была как можно меньше. Перед нами – парабола, ветки которой направлены вверх. Тогда неплохо бы смотреть на вершину, а значит

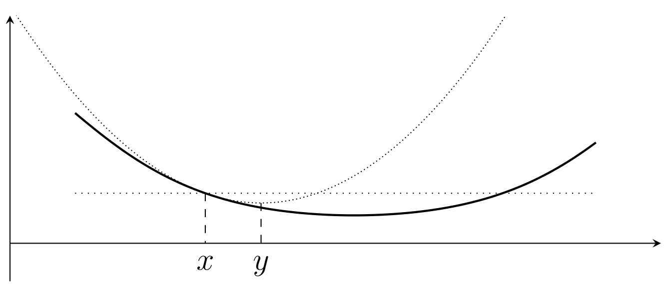
$$\alpha = \frac{1}{L}.$$

Подставив, получим оценку сверху на уменьшение функции:

$$f(y) \leqslant f(x) - \frac{1}{2L}\|\nabla f(x)\|^2.$$

Тогда величина $\frac{1}{2L}\|\nabla f(x)\|^2$ показывает гарантированную величину, на которую уменьшится значение функции на этом шаге.

Замечание. *На рисунке представлена парабола, ограничивающая нашу функцию сверху. Мы, по сути, ищем минимум этой параболы. Кроме того видно, что шаг больший, чем $\frac{1}{L}$ брать точно не нужно: гарантий, что значение на следующей итерации уменьшится, вовсе нет. Это общая ситуация для методов спуска: шаг не должен быть слишком маленьким, чтобы на шаге был ощутимый прогресс и алгоритм работал достаточно быстро. В то же время, шаг не должен быть достаточно большим, ведь тогда нет гарантии хоть какого-то прогресса на каждом следующем шаге.*



Продолжим наш анализ. Мы получаем, что

$$f(x_{k+1}) \leqslant f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \leqslant f(x_{k-1}) - \frac{1}{2L}\|\nabla f(x_k)\|^2 - \frac{1}{2L}\|\nabla f(x_{k-1})\|^2 \leqslant$$

$$\leqslant \ldots \leqslant f(x_0) - \frac{1}{2L}\sum_{i=0}^{k}\|\nabla f(x_i)\|^2.$$

Перегруппировав, получим

$$\sum_{i=0}^{k} \|\nabla f(x_i)\|^2 \leqslant 2L(f(x_0) - f(x_{k+1})).$$

Предполагаем, что $f(x) > -\infty$, а тогда $f(x_{k+1}) \geqslant f_{opt}$, значит

$$\sum_{i=0}^{k} \|\nabla f(x_i)\|^2 \leqslant 2L(f(x_0) - f_{opt}).$$

Пусть $g_k = \min_{0 \leqslant i \leqslant k} \|\nabla f(x_i)\|^2$, тогда

$$(k+1)g_k \leqslant \sum_{i=0}^{k} \|\nabla f(x_i)\|^2 \leqslant 2L(f(x_0) - f_{opt}) \ \Rightarrow \ g_k \leqslant \frac{2L(f(x_0) - f_{opt})}{k+1}.$$

Видно, что сходимость здесь сублинейная. Итого, справедлива следующая теорема.

Теорема 4. *Пусть $f \in C_L^{1,1}$, $f \geqslant f_{opt}$, $\alpha = \frac{1}{L}$. Тогда в градиентном спуске справедлива следующая оценка:*

$$g_k \leqslant \frac{2L(f - f_{opt})}{k+1}.$$

Замечание. *Никто не гарантирует, что метод сойдется к минимуму. Метод может сойтись к седловой точке.*

Сходимость это всегда хорошо, но сублинейная скорость сходимости это довольно слабо. Можно быть, бывают случаи, когда градиентный спуск работает эффективнее? Оказывается, да. Но прежде чем перейти к этому случаю, давайте введем некоторые дополнительные определения.

3.4 Выпуклые и μ-сильно выпуклые функции

Напомним определение выпуклой функции:

Определение 11. *Пусть $f : X \to \mathbb{R}$, X – выпуклое множество. Если*

$$f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha)f(y), \quad x, y \in X, \quad \alpha \in [0,1],$$

то f называется выпуклой вниз на X. Если неравенство оказывается строгим при $x \neq y$, $\alpha \in (0,1)$, то функция называется строго выпуклой.

Для дифференцируемых функций можно дать другую характристику, с геометрической интерпретацией «функция выпукла вниз, если она лежит выше касательной».

Теорема 5. *Пусть f – дифференцируемая на выпуклом множестве $X \subset \mathbb{R}^n$ функция. Тогда f выпукла вниз тогда и только тогда, когда*

$$f(y) \geqslant f(x) + f'(x)^T(y-x)$$

при $x, y \in X$.

Доказательство. Из определения выпуклости при $\alpha \in (0, 1]$ имеем

$$f(\alpha y + (1-\alpha)x) \leqslant \alpha f(y) + (1-\alpha) f(x),$$

откуда

$$f(y) - f(x) \geqslant \frac{f(\alpha y + (1-\alpha)x) - f(x)}{\alpha} = \frac{f(x + \alpha(y-x)) - f(x)}{\alpha} =$$
$$= \frac{f'(x)^T \alpha(y-x) + o(\alpha)}{\alpha} = f'(x)^T(y-x).$$

Обратно, пусть

$$f(y) \geqslant f(x) + f'(x)^T(y-x).$$

Пусть $x = \alpha x_1 + (1-\alpha)x_2$. Тогда

$$f(x_1) \geqslant f(x) + f'(x)^T(x_1 - x), \quad f(x_2) \geqslant f(x) + f'(x)^T(x_2 - x).$$

Значит,

$$\alpha f(x_1) + (1-\alpha) f(x_2) \geqslant \alpha f(x) + (1-\alpha) f(x) + \alpha f'(x)^T((1-\alpha)x_1 - (1-\alpha)x_2) +$$
$$+ (1-\alpha) f'(x)^T(-\alpha x_1 + \alpha x_2) = f(x).$$

□

Понятно, что для строгой выпуклости можно требовать выполнения строгого неравенства при $x \neq y$.

Определение 12. *Функция f называется сильно выпуклой вниз с коэффициентом μ, если выпукла вниз функция*

$$f(x) - \frac{\mu}{2}\|x\|^2.$$

Другими словами, мы требуем, чтобы выпуклость для функции была в каком-то смысле сильнее, чем у хоть какой-нибудь «параболы».

Определение 13. *Пусть $f : X \to \mathbb{R}$, X – выпуклое множество. Функция f называется сильно выпуклой вниз с коэффициентом $\mu > 0$, если*

$$f(\alpha x + (1 - \alpha y) \leqslant \alpha f(x) + (1-\alpha) f(y) - \frac{\mu}{2}\alpha(1-\alpha)\|x-y\|^2,$$

$x, y \in X$, $\alpha \in [0, 1]$

Лемма 1. *Приведенные определения эквивалентны.*

Доказательство. Докажем необходимость. Используя определения выпуклости, имеем

$$f(\alpha x + (1-\alpha)y) - \frac{\mu}{2}\|\alpha x + (1-\alpha)y\|^2 \leqslant$$
$$\leqslant \alpha f(x) + (1-\alpha) f(y) - \alpha\frac{\mu}{2}\|x\|^2 - (1-\alpha)\frac{\mu}{2}\|y\|^2.$$

Так как любая норма – выпуклая вниз функция, то

$$\|\alpha x + (1-\alpha)y\|^2 \leqslant (\alpha\|x\| + (1-\alpha)\|y\|)^2.$$

Раскрывая скобки и приводя подобные, приходим к требуемому. Обратное утверждение доказывается обратными преобразованиями. □

Покажем геометрическую интерпретацию теоремы, годную для дифференцируемой функции.

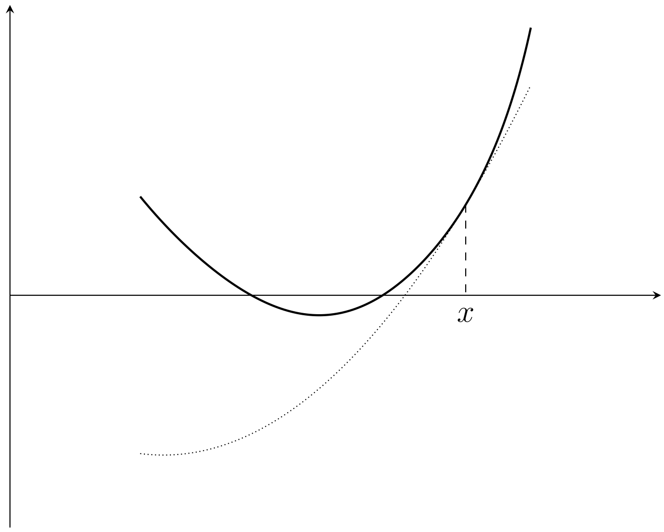
Теорема 6. *Пусть $f : X \to \mathbb{R}$, X – выпуклое множество, f дифференцируема на X. f сильно выпукла вниз тогда и только тогда, когда $\forall x \in X$ неравенство*

$$f(y) \geqslant f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|x - y\|^2$$

выполнено при всех $y \in f$.

Доказательство. Данная теорема доказывается точно так же, как аналогичная теорема для выпуклых функций и их характеристики в терминах касательной гиперплоскости. □

Итого, важно следующее: сильно выпуклая функция в каждой точке может быть «подперта» снизу параболоидом с положительным коэффициентом при квадрате.



Замечание. *В случае, когда $f \in C^2(\mathbb{R}^n)$, то $\mu = \inf\limits_{x \in \mathbb{R}^n} \lambda_{\min}(\nabla^2 f(x))$*

Замечание. *Очевидно, что любая сильно выпуклая функция является и выпуклой. Можно вообще сказать, что выпуклая функция – это сильно выпуклая функция с коэффициентом $\mu = 0$. Обратное, вообще говоря, неверно. Например, функция $y = x^4$ является выпуклой вниз, но не сильно выпуклой.*

3.5 Градиентный спуск для $C_L^{1,1}$ и μ-сильно выпуклых функций

Пусть $f \in C_L^{1,1}$ и μ-сильно-выпукла. Тогда справедлива сразу пара неравенств:

$$f(y) \leqslant f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|^2,$$

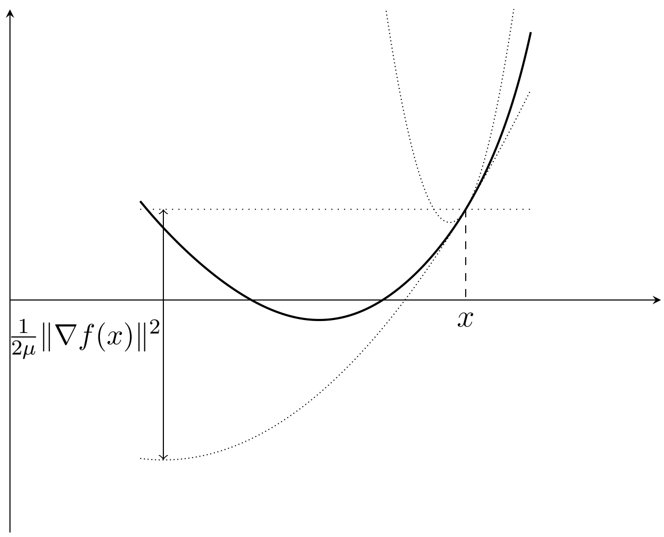
$$f(y) \geqslant f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2.$$

Обозначим правую часть последнего неравенства через $h(y)$, считая, что x фиксирован. Имеем

$$f(y) \geqslant h(y), \quad x, y \in \mathbb{R}^n.$$

Отсюда следует, что
$$f_{opt} = \min_y f(y) \leqslant \min_y h(y).$$
Минимизируем нашу функцию $h(y)$. Возьмем градиент по y:
$$\nabla h(y) = \nabla f(x) + \mu(y - x).$$
Из условий оптимальности получим, что
$$y_{opt} = x - \frac{1}{\mu}\nabla f(x) \Rightarrow \; h(y_{opt}) = f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2.$$
Итого,
$$f_{opt} \geqslant f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2$$

Замечание. *А что это означает геометрически?*



На рисунке представлена парабола, ограничивающая нашу функцию снизу (за счет сильной выпуклости). Видно, что разница между текущим и оптимальным значением оценивается как раз-таки через $\frac{1}{2\mu}\|f(x)\|^2$. *Понятно, что если у нас есть лишь информация о том, что функция выпукла, то снизу оценить функцию мы можешь лишь касательной, уходящей в бесконечность, а значит оценку на невязку по функции таким простым способом, без дополнительных предположений и инструментов, не получить.*

Теперь получим оценку на невязку по функции
$$f_{opt} \geqslant f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2 \;\Leftrightarrow\; 2\mu(f(x) - f_{opt}) \leqslant \|\nabla f(x)\|^2.$$
Тогда
$$f(x_{k+1}) - f_{opt} \leqslant f(x_k) - f_{opt} - \frac{1}{2L}\|\nabla f(x_k)\|^2 \leqslant f(x_k) - f_{opt} - \frac{\mu}{L}(f(x_k) - f_{opt}) =$$
$$= \left(1 - \frac{\mu}{L}\right)(f(x_k) - f_{opt}).$$

Итого,

$$f(x_{k+1}) - f_{opt} \leqslant \left(1 - \frac{\mu}{L}\right)^{k+1} (f(x_0) - f_{opt}).$$

Видно, что теперь сходимость оказалась линейной. Итого, справедлива следующая теорема.

Теорема 7. *Пусть $f \in C_L^{1,1}(\mathbb{R}^n)$, причем f – μ-сильно-выпукла, $\alpha = \frac{1}{L}$. Тогда в GD имеем:*

$$f(x_k) - f_{opt} \leqslant \left(1 - \frac{\mu}{L}\right)^{k} (f(x_0) - f_{opt}).$$

3.6 Адаптивный поиск константы Липшица

К сожалению, для прикладных задач не всегда получается найти константу Липшица. В таком случае можно воспользоваться следующим эмпирическим алгоритмом, который построен на уже известном нам неравенстве:

$$f(x_{k+1}) \leqslant f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2.$$

Данный алгоритм называют адаптивным поиском или адаптивной стратегией.

Алгоритм 5 Стратегия выбора L

Вход: Функция f, начальное приближение L_0, x_0 и так далее.

Повторять пока не вышли **Выполнять**
    $x_{k+1} = x_k - \frac{1}{L_k}\nabla f(x_k)$
    Если $f(x_{k+1}) \leqslant f(x_k) - \frac{1}{2L_k}\|\nabla f(x_k)\|^2$ **тогда**
    Конец условия
    $L_k := L_k\gamma$, $\gamma > 1$.
Конец цикла
$L_{k+1} = L_k\rho$, $\rho < 1$.

Параметры γ и ρ остаются на откуп применяющего.

Замечание. *За счёт адаптивности, метод может работать даже быстрее, чем при истинной константе Липшица. Ведь на самом деле нас интересуют свойства функции скорее локально, в некоторой окрестности нашего текущего шага, нежели глобально, на всём $\mathbb{R}^n$.*

3.7 Стратегия Армихо-Вульфа

Зададимся вопросом «А что вообще такое спуск»? Кажется, логично, что это движение в направлении уменьшения. Сложно ввести такую характеристику глобально, но для локального уменьшения у нас есть способ.

Определение 14. *Пусть f - дифференцируемая функция. Направление d будем называть направлением спуска в точке x, если*

$$\nabla f(x)^T d < 0$$

Замечание. *Абсолютно логично, что градиентный спуск является спуском:*

$$f(x)^T(-\nabla f(x)) = -\|\nabla f(x)\|^2 < 0$$

Рассмотрим теперь некоторую общую стратегию. Как один из вариантов подбора шага, возможно рассмотреть решение следующей задачи:

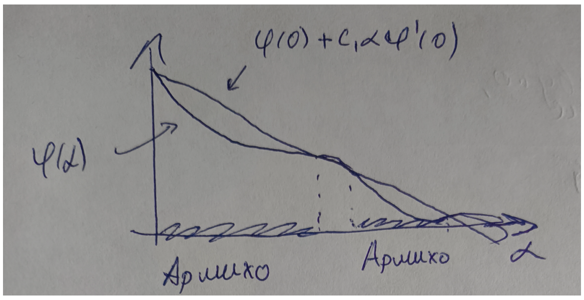
$$\varphi(\alpha) = f(x_k - \alpha d) \to \min_{\alpha > 0}$$

Если искать честное решение (может быть лишь с ограниченным диапазоном поиска α), получится что-то близкое к методу наискорейшего спуска. Но на самом деле нет необходимости искать значение точно на каждой итерации, ведь итераций много. Тут и возникает понятие неточной оптимизации, когда мы зададимся некоторым удачным набором условий. Но что считать удачным?

Если d - направление спуска, очевидно, что $\varphi'(0) = -\nabla f(x_k)^T d < 0$

Определение 15. *Будем говорить, что α удовлетворяет условиям Армихо, если*

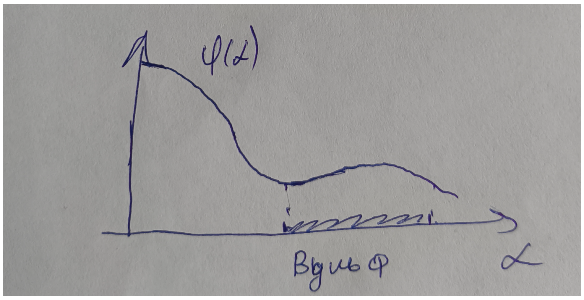
$$\varphi(\alpha) < \varphi(0) + c_1 \alpha \varphi'(0), \quad c_1 \in (0,1)$$



Означает написанное, по сути, что мы можем отсечь некоторой стороной угла нашу функцию так, что будут видны подходящие α, см. рисунок. Нас интересует такое достаточное α, чтобы функция значительно уменьшилась на принятом шаге.

Самого этого условия недостаточно: ведь можно брать достаточно маленькие значение α. Понятно, что надо брать какие-то компромиссные шаги.

Определение 16. *Говорят, что функция удовлетворяет слабому условию Вульфа, если $\varphi'(\alpha) \geqslant c_2\varphi'(0)$. Говорят, что функция удовлетворяет сильному условию Вульфа, если $|\varphi'(\alpha)| \leqslant c_2|\varphi'(0)|$, $c_2 \in (0,1)$.*



Слабое условие, по сути, ограничивает производную снизу. Сильную – с двух сторон. Выбором коэффициента c_2 мы можем стараться приближение неточного минимума стремить к точному.

Справделивым будет спросить, а правда ли найдутся такие значения? Следующая лемма, которую мы приведём без доказательств, отвечает на него теоретически.

Лемма 2. *Пусть $\varphi \in C^1$, $\varphi'(0) < 0$, $\varphi(\alpha) > -\infty$, $0 < C_1 \leqslant C_2 < 1$. Тогда найдется точка α, удовлетворяющая условиям Армихо и Вульфа.*

На практике же мы предложим следующий алгоритм, который в большей степени расчитан на условие Армихо, чтобы гарантированно сделать шаг, но может быть дополнен и до условия Вульфа.

Алгоритм 6 backtracking

Вход: α.

Инициализация:

$$\alpha = \alpha_{start}$$

Если $\varphi(\alpha) \leqslant \varphi(0) + c_1\alpha\varphi'(0)$, **тогда**

    стоп

иначе

    $\alpha := \alpha\rho$, $\rho \in (0, 1)$.

Конец условия

3.8 Градиентный спуск для $C_L^{1,1}$ со стратегией Армихо-Вульфа

Давайте подбирать величину шага с использованием стратегии Армихо и Вульфа. Согласно стратегии Армихо, при неточной одномерной минимизации функции

$$\varphi(\alpha) = f(x_k - \alpha\nabla f(x_k)), \quad \alpha > 0,$$

нам подходят те значения α, которые удовлетворяют условию

$$\varphi(\alpha) \leqslant \varphi(0) + c_1\alpha\varphi'(0), \quad c_1 \in (0, 1).$$

Понятно, что в нашем случае

$$\varphi(0) = f(x_k), \quad \varphi'(0) = -\|\nabla f(x_k)\|^2 < 0.$$

Согласно (слабой) стратегии Вульфа, чтобы не допускать слишком маленьких по величине шагов α, мы требуем выполнения и следующего неравенства:

$$\varphi'(\alpha) \geqslant c_2\varphi'(0), \quad \varphi'(\alpha) = -\nabla f(x_{k+1})^T\nabla f(x_k).$$

Стратегия Армихо уже дает нам нужное для сходимости градиентного спуска неравенство:

$$f(x_{k+1}) \leqslant f(x_k) - c_1\alpha\|\nabla f(x_k)\|^2.$$

Покажем, что стратегия Вульфа отделяет написанную константу от нуля. Имеем,

$$-\nabla f(x_{k+1})^T\nabla f(x_k) \geqslant -c_2\|\nabla f(x_k)\|^2.$$

Рассмотрим следующее выражение:

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

$$\leqslant L\|x_k - x_{k+1}\| \|\nabla f(x_k)\| = L\alpha \|\nabla f(x_k)\|^2.$$

С другой стороны,

$$(\nabla f(x_k) - \nabla f(x_{k+1}))^T \nabla f(x_k) = \|\nabla f(x_k)\|^2 - \nabla f(x_{k+1})^T \nabla f(x_k) \geqslant$$

$$\geqslant \|\nabla f(x_k)\|^2 - c_2\|\nabla f(x_k)\|^2 = (1 - c_2)\|\nabla f(x_k)\|^2,$$

откуда

$$\alpha \geqslant \frac{(1 - c_2)}{L}.$$

В итоге,

$$f(x_{k+1}) \leqslant f(x_k) - c_1\alpha\|\nabla f(x_k)\|^2 \leqslant f(x_k) - \frac{c_1(1 - c_2)}{L}\|\nabla f(x_k)\|^2,$$

что и гарантирует отделимость α от нуля, а значит гарантированное уменьшение значения функции на каждом шаге. Результаты про сходимость качественно не меняются (хотя и меняются константы).

4 Метод Ньютона

4.1 Эмпирический подход к методу Ньютона

Продолжаем решать следующую задачу:

$$f(x) \to \min, \quad x \in \mathbb{R}^n.$$

Поменяем модель приближения функции. Будем рассматривать квадратичную функцию, то есть модуль следующего вида

$$f(x+h) \approx f(x) + \nabla f(x)^T h + \frac{1}{2} h^T B h = g(h)$$

Будем предполагать, что $B = B^T$, B положительно определена. По сути, это означает, что f неплохо приближается выпуклой g. Найдем минимум g для определения оптимального шага h:

$$\nabla g(h) = \nabla f(x) + Bh = 0 \quad \Rightarrow \quad h = -B^{-1}\nabla f(x).$$

Замечание. *Полезно ответить, что в случае, когда $B = I$ – единичная матрица, то перед нами снова метод градиентного спуска. Итак, другой подход к выводу и пониманию метода градиентного спуска – аппроксимация параболоидом с одинаковой кривизной по всем направлениям. Понятно теперь, почему число обусловленности важно.*

Разумнее, конечно, рассмотреть в качестве $B = \nabla^2 f(x)$ – гессиан. Для выпуклой вниз функции это возможно в силу его положительной определенности. В этом случае

$$h = -(\nabla^2 f(x))^{-1}\nabla f(x).$$

Написанное корректно: положительно определенная симметричная матрица обратима. Итак, метод Ньютона работает по следующей схеме выбора новой точки:

$$x_{k+1} = x_k - \alpha(\nabla^2 f(x_k))^{-1}\nabla f(x_k).$$

Во многих источниках говорится, что $\alpha = 1$. Это неприятная неточность. О ней мы поговорим позже. Построенный метод и правда является методом спуска, ведь

$$\nabla f(x)^T(\nabla^2 f(x))^{-1}\nabla f(x) > 0$$

при $\nabla f(x) \neq 0$ согласно предположению о положительной определенности. Понятно, что построенный метод решает задачу минимизации квадратичной функции просто за одну итерацию – получаем так называемые многомерный метод парабол. Более того, метод инвариантен к обратимым линейным преобразованиям переменных.

Лемма 3. *Пусть решается задача*

$$f(x) \to \min.$$

Если $x = Ay$, $\det A \neq 0$, то исходная задача и ее решение равносильны задаче

$$r(y) = f(Ay) \to \min, \quad \widetilde{x} \in \mathbb{R}^n.$$

Доказательство. Найдём градиент и гессиан функции $r(y)$:

$$\nabla r(y) = A^T f(Ay), \qquad \nabla^2 r(y) = A^T f(Ay) A.$$

Теперь оценим, как происходит шаг от $x_k = Ay_k$ к x_{k+1}

$$x_{k+1} = x_k - \alpha_k [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)) = Ay_k - \alpha_k [\nabla^2 f(Ay_k)]^{-1} \nabla f(Ay_k) =$$
$$Ay_k - \alpha_k AA^{-1} [\nabla^2 f(Ay_k)]^{-1} (A^T)^{-1} A^T \nabla f(Ay_k) = Ay_k - A\alpha_k [A^T \nabla^2 f(Ay_k) A]^{-1} A^T \nabla f(Ay_k) =$$
$$= A\left(y_k - \alpha_k [\nabla^2 r(y_k)]^{-1} \nabla r(y_k)\right) = Ay_{k+1}$$

□

4.2 Глобальная сходимость метода Ньютона

Сначала поговорим о глобальной сходимости метода Ньютона. Для этого несколько в ином виде перепишем разговоры, проводимые ранее про стратегию Армихо и Вульфа.

Итак, в общем случае метод спуска выглядит следующим образом:

$$x_{k+1} = x_k + \alpha_k d_k,$$

где d_k – направление убывания функции, то есть $\nabla f(x_k)^T d_k < 0$. Значение шага α_k выбираем, используя стратегию Армихо и Вульфа (например, слабую). Напомним, что речь идет об одномерной неточной минимизации функции

$$\varphi(\alpha) = f(x_k + \alpha d_k), \quad \varphi'(0) = \nabla f(x_k)^T d_k < 0.$$

Условия Армихо говорят, что нам подходят те точки, для которых

$$\varphi(\alpha) \leqslant \varphi(0) + c_1 \alpha \varphi'(0).$$

В терминах функции, это переписывается в виде

$$f(x_{k+1}) \leqslant f(x_k) + c_1 \alpha \nabla f(x_{k+1})^T d_k.$$

Стратегия Вульфа определяет и следующие границы на α:

$$\varphi'(\alpha) \geqslant c_2 \varphi'(0).$$

В терминах функций мы можем переписать все в виде

$$\nabla f(x_{k+1})^T d_k \geqslant c_2 \nabla f(x_k)^T d_k.$$

Итого, выполнение обоих условий одновременно означает, что

$$\begin{cases} f(x_{k+1}) \leqslant f(x_k) + c_1 \alpha \nabla f(x_{k+1})^T d_k \\ \nabla f(x_{k+1})^T d_k \geqslant c_2 \nabla f(x_k)^T d_k \end{cases}.$$

Покажем, что выполнение этих условий уменьшает значение функции на следующей итерации. Заметим, что это уже гарантирует первое условие и условие на выбор d_k. В то же время, чтобы исключить возможность маленьких шагов, которые могут не отделять от нуля последнее выражение первого неравенства, воспользуемся вторым неравенством. Как обычно, рассматриваем функции $f \in C_L^{1,1}$.

$$\left|(\nabla f(x_{k+1}) - \nabla f(x_k))^T d_k\right| \leqslant \|\nabla f(x_{k+1}) - f(x_k)\|\|d_k\| \leqslant L\|x_{k+1} - x_k\|\|d_k\|.$$

Так как $x_{k+1} = x_k + \alpha d_k$, то окончательно

$$\left|(\nabla f(x_{k+1}) - \nabla f(x_k))^T d_k\right| \leqslant L\alpha\|d_k\|^2.$$

С другой стороны, используя условие Вульфа,

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^T d_k = \nabla f(x_{k+1})^T d_k - \nabla f(x_k)^T d_k \geqslant (c_2 - 1)\nabla f(x_k)^T d_k,$$

откуда

$$(c_2 - 1)\nabla f(x_k)^T d_k \leqslant L\alpha\|d_k\|^2 \;\Rightarrow\; \alpha \geqslant \frac{(c_2 - 1)}{L\|d_k\|^2}\nabla f(x_k)^T d_k.$$

Теперь воспользуемся первым неравенством, тогда

$$f(x_{k+1}) \leqslant f(x_k) + c_1\alpha\nabla f(x_{k+1})^T d_k \leqslant f(x_k) - \frac{c_1(1 - c_2)}{L}\frac{(\nabla f(x_k)^T d_k)^2}{\|d_k\|^2}.$$

Немного переобозначим все написанное. Ясно, что

$$\nabla f(x)^T d_k = \|\nabla f(x)\|\|d_k\|\cos\theta_k,$$

откуда неравенство можно переписать в виде

$$f(x_{k+1}) \leqslant f(x_k) - \frac{c_1(1 - c_2)}{L}\|\nabla f(x_k)\|^2\cos^2\theta_k.$$

В нашем случае

$$d_k = -(\nabla^2 f(x_k))^{-1}\nabla f(x_k),$$

тогда

$$\cos\theta_k = -\frac{\nabla f(x_k)^T(\nabla^2 f(x_k))^{-1}\nabla f(x_k)}{\|\nabla f(x_k)\|\|(\nabla^2 f(x_k))^{-1}\nabla f(x_k)\|}.$$

Стандартно имеем:

$$\lambda_{\min}(A)\|x\|^2 \leqslant x^T A x \leqslant \lambda_{\max}(A)\|x\|^2.$$

Так как

$$\|(\nabla^2 f(x_k))^{-1}\nabla f(x_k)\| = \sqrt{\|\nabla f(x_k)^T(\nabla^2 f(x_k))^{-2}\nabla f(x_k)\|} \leqslant \sqrt{\lambda_{\max}((\nabla^2 f(x_k))^{-2})} =$$

$$= \lambda_{\max}((\nabla^2 f(x_k))^{-1}),$$

то

$$\cos\theta_k \leqslant -\frac{\lambda_{\min}(\nabla^2 f(x_k))^{-1}\|\nabla f(x_k)\|^2}{\lambda_{\max}(\nabla^2 f(x_k))^{-1}\|\nabla f(x_k)\|^2} = -\frac{\lambda_{\min}(\nabla^2 f(x_k))^{-1}}{\lambda_{\max}(\nabla^2 f(x_k))^{-1}}$$

Это гарантирует отделимость косинуса от нуля и глобальную сходимость, как и в случае метода градиентного спуска.

4.3 Локальная суперлинейная (квадратичная) сходимость

Теорема 8. *Пусть* $f \in C_M^{2,2}$, $\nabla^2 f(x_{opt}) \geqslant \mu I$, $\mu > 0$, $\alpha_k = 1$. *Тогда существует* $r > 0$, *что*

$$\forall x: \ \|x - x_{opt}\| < r$$

для метода Ньютона выполнено

$$\|x_{k+1} - x_{opt}\| \leqslant \|x_k - x_{opt}\|^2.$$

Доказательство. Пусть $H_k = \nabla^2 f(x_k)$. Рассмотрим следующие преобразования:

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

$$= \|H_k^{-1}(H_k(x_k - x_{opt}) - (\nabla f(x_k) - \nabla f(x_{opt}))\| \leqslant$$

$$\|H_k^{-1}\| \left\| H_k(x_k - x_{opt}) - \int\limits_0^1 \nabla^2 f(x_{opt} + \tau(x_k - x_{opt}))(x_k - x_{opt})d\tau \right\| =$$

$$= \|H_k^{-1}\| \left\| \int\limits_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x_{opt} + \tau(x_k - x_{opt})\right)(x_k - x_{opt})d\tau \right\| \leqslant$$

$$\leqslant \|H_k^{-1}\| M \int\limits_0^1 (1-\tau)d\tau \|x_k - x_{opt}\|^2 \leqslant \frac{M}{\mu}\|x_k - x_{opt}\|^2.$$

□

4.4 Метод Ньютона для невыпуклых функций

Модификация, вроде бы, понятна:

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

где $B_k = \nabla^2 f(x_k) + E_k$, B_k положительно определена. Осталось понять, как определить и модифицировать B_k путем выбора E_k.

1. Первый способ – модификация собственного разложения.

   $$\nabla^2 f(x_k) = Q\Lambda Q^T,$$

   где Q – ортонормированная матрица из собственных векторов. Пусть

   $$B_k = Q\widetilde{\Lambda}Q^T, \quad \widetilde{\lambda}_i = \begin{cases} \lambda_i, & \lambda_i \geqslant \delta > 0 \\ \delta, & \text{иначе} \end{cases}.$$

   Для определения B_k^{-1} сразу пользуемся имеющимся разложением, решая соответствующую систему уравнений:

   $$B_k d = -\nabla f(x_k) \ \Leftrightarrow \ Q\widetilde{\Lambda}Q^T d = -\nabla f(x_k) \ \Rightarrow \ d = -Q\widetilde{\Lambda}Q^T\nabla f(x_k).$$

   Недостаток, собственное разложение – дорогое разложение.

2. Второй способ – модификация диагонали матрицы гессиана.

$$B_k = \nabla^2 f(x_k) + \tau_k I.$$

Перед нами – просто-напросто сдвиг всех собственных значений на τ_k, ведь

$$I = QQ^T \;\Rightarrow\; B_k = Q(\Lambda + \tau_k I)Q^T.$$

Приложения

Численная производная и машинная точность

Почему нельзя взять, и уменьшать шаг производной до самого малого?