Задание 10. Методы безусловной многомерной оптимизации

Спонсор задания — В. И. Гориховский

28 марта 2025

[1] Б. Т. Поляк. Введение в оптимизацию.

[2] В. Г. Жадан. Методы оптимизации. Часть 2: численные алгоритмы.

[3] Ссылки:
http://mech.math.msu.su/~vvb/MasterAI/GradientDescent.html
https://pnu.edu.ru/media/filer_public/2013/02/26/popova_methods-mo.pdf
https://arxiv.org/ftp/arxiv/papers/1711/1711.00394.pdf
https://www.mathnet.ru/links/f8f8999f792f184fa8cdf0721137df7f/at554.pdf

[4] П. К. Силаев, В. А. Ильина. Численные методы для физиков-теоретиков. Часть I. 2003.

Что нужно сделать:

Реализовать и сравнить методы по метрикам:

- скорость сходимости по числу вычислений оптимизируемой функции
- скорость сходимости по времени
- точность метода
- вероятность нахождения глобального оптимума

Лирическое

Методы поиска: безградиентные (нулевого порядка) и градиентные (первого порядка; второго...)

Совет от физиков-теоретиков: если нет возможности вычислить градиент хотя бы приблизительно, НЕ стоит применять градиентные методы, вычисляя градиент численно покоординатным варьированием.

Метод градиентного спуска — не всегда разумный способ найти многомерный минимум.

Градиентный метод

Считаем, что в любой точке x можем вычислить градиент функции $\nabla f(x)$.
Тогда, начиная с некоторого приближения x^0, строим итерационную последовательность

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k), \text{ где } \gamma_k \geqslant 0 \text{ — длина шага.}$$

Как прийти к этому методу?

1. Из необходимых условий экстремума: если в точке x условие экстремума не выполняется ($\nabla f(x) \neq 0$), то значение функции можно уменьшить, перейдя к точке $x - \tau \nabla f(x)$ при достаточно малом $\tau > 0$. Применяем эту идею итеративно.
2. В точке x^k дифференцируемая функция $f(x)$ приближается линейной $f_k(x) = f(x^k) + (\nabla f(x^k), x - x^k)$ с точностью до членов порядка $o(x - x^k)$. Поэтому можно искать минимум аппроксимации $f_k(x)$ в окрестности x^k.
   - Например, можно задать некоторое ε_k и решать вспомогательную задачу $\min_{\|x - x^k\| \leqslant \varepsilon_k} f_k(x)$. Решение этой задачи принимают за новое приближение x^{k+1}.
   - Можно остаться в окрестности x^k иначе: добавив к $f_k(x)$ «штраф» за отклонение от x^k. Например, решая вспомогательную задачу $\min\left[f_k(x) + \alpha_k \|x - x^k\|^2\right]$ и ее решение брать в качестве x^{k+1}.
3. Можно в точке x^k брать направление локального наискорейшего спуска (оно будет противоположно направлению градиента).

Сходимость градиентного метода

Считаем, что $\gamma_k \equiv \gamma$. То есть метод таков: $x^{k+1} = x^k - \gamma \nabla f(x^k)$.

Theorem

Пусть $f(x)$ дифференцируема на $\mathbb{R}^n$,

- *градиент $f(x)$ удовлетворяет условию Липшица: $\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|$,*
- *$f(x)$ ограничена снизу: $f(x) \geqslant f^* > -\infty$,*
- *и γ удовлетворяет условию $0 < \gamma < 2/L$.*

Тогда в методе

- *градиент стремится к нулю: $\lim_{k\to\infty} \nabla f(x^k) = 0$,*
- *а функция $f(x)$ монотонно убывает: $f\left(x^{k+1}\right) \leqslant f\left(x^k\right)$.*

Сходимость для сильно выпуклых функций

Функция $f(x)$ на $\mathbb{R}^n$ называется сильно выпуклой с константой $\ell > 0$, если при $0 \leqslant \lambda \leqslant 1$

$$f\big(\lambda x + (1-\lambda)y\big) \leqslant \lambda f(x) + (1-\lambda)f(y) - \ell\lambda(1-\lambda)\|x-y\|^2/2.$$

Theorem

Пусть $f(x)$ дифференцируема на $\mathbb{R}^n$, ее градиент удовлетворяет условию Липшица с константой L и $f(x)$ является сильно выпуклой функцией с константой $\ell > 0$.
Тогда при $0 < \gamma < 2/L$ метод сходится к единственной точке глобального минимума x^* со скоростью геометрической прогрессии: $\|x^k - x^*\| \leqslant cq^k$, $0 \leqslant q < 1$.

Метод Ньютона

Используем квадратичную (а не линейную) аппроксимацию функции в точке x^k, т.е. функцию

$$f_k(x) = f(x^k) + \left(\nabla f(x^k), x - x^k\right) + \left(\nabla^2 f(x^k) \cdot (x - x^k), x - x^k\right)/2.$$

Выбираем точку минимума $f_k(x)$ в качестве нового приближения: $x^{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} f_k(x)$.

Сам метод: $x^{k+1} = x^k - \left[\nabla^2 f(x^k)\right]^{-1} \cdot \nabla f(x^k)$.

Как придумать метод чуть по-другому?
Точка минимума должна быть решением системы n уравнений с n переменными: $\nabla f(x) = 0$, каковую решают методом Ньютона, заключающимся в линеаризации уравнений в точке x^k и решении линеаризованной системы.

Сходимость метода Ньютона

Theorem

Пусть $f(x)$ дважды дифференцируема,

- *$\nabla^2 f(x)$ удовлетворяет условию Липшица с константой L,*
- *$f(x)$ сильно выпукла с константой ℓ*
- *и начальное приближение удовлетворяет условию $q = \frac{L}{2\ell^2}\|\nabla f(x^0)\| < 1$.*

Тогда метод сходится к точке глобального минимума x^* с квадратичной скоростью: $\|x^k - x^*\| \leqslant \frac{2\ell}{L} q^{2^k}$.

Выводы

Метод	«+»	«-»
Градиентный	Глобальная сходимость. Слабые требования к $f(x)$. Простота вычислений.	Медленная сходимость. Необходимость выбора γ.
Ньютона	Быстрая сходимость	Локальная сходимость. Жесткие трабования к $f(x)$. Большой объем вычислений.

Модификации градиентного метода

Идея: различные способы выбора длины шага γ_k (см. Приложение А).

1. Идти до минимума по направлению антиградиента:

$$\gamma_k = \text{argmin}_{\gamma \geqslant 0} \varphi_k(\gamma), \ \varphi_k(\gamma) = f\left(x^k - \gamma \nabla f(x^k)\right).$$

Получаем *метод скорейшего спуска.*

Про сходимость и скорость сходимости доказано в [1].

Но, например, для квадратичной функции метод скорейшего спуска сходится не быстрее, чем простой градиентный метод при соответствующем выборе γ.

2. Выбор $\gamma_k \equiv \gamma$, $0 < \gamma < 2/L$ — некоструктивен из-за незнания L.

Но можно использовать процедуру подбора γ:

задают $0 < \varepsilon < 1$, $0 < \alpha < 1$ и некоторое γ. На каждой итерации вычисляют $\xi = f\left(x^k - \gamma \nabla f(x^k)\right)$ и проверяют неравенство $\xi \leqslant f(x^k) - \varepsilon\gamma \|\nabla f(x^k)\|^2$.

Если это верно, то $x^{k+1} := x^k - \gamma \nabla f(x^k)$, иначе $\gamma := \alpha\gamma$ и проверку повторяют.

Первая модификации метода Ньютона: демпфированный метод Ньютона

Первая идея: регулирем длину шага, т.е.

$$x^{k+1} = x^k - \gamma_k \left[\nabla^2 f(x^k)\right]^{-1} \cdot \nabla f(x^k)$$

— демпфированный метод Ньютона.

- $\gamma_k = \text{argmin}_{\gamma \geqslant 0} f\left(x^k - \gamma \left[\nabla^2 f(x^k)\right]^{-1} \cdot \nabla f(x^k)\right)$
- или γ дробится (умножается на $0 < \alpha < 1$), начиная с $\gamma = 1$, до выполнения какого-либо условия

$$f(x^{k+1}) \leqslant f(x^k) - \gamma q \left(\left[\nabla^2 f(x^k)\right]^{-1} \cdot \nabla f(x^k), \nabla f(x^k)\right), \ 0 < q < 1,$$

$$\text{или } \|\nabla f(x^{k+1})\|^2 \leqslant (1 - \gamma q)\|\nabla f(x^k)\|^2, \ 0 < q < 1.$$

Для гладких сильно выпуклых функций метод глобально сходится.
На начальных итерациях сходимость со скоростью геометрической прогрессии.
При попадании в окрестность x^* будет иметь место квадратичная сходимость.

Вторая модификация: метод Левенберга–Марквардта

Направление движения отличается от задаваемого методом Ньютона: к аппроксимирующей функции добавляется квадратичный штраф за отклонение от точки x^k.
Т.е. ищут x^{k+1} из условия минимума

$$f_k(x) + \frac{\alpha_k}{2}\|x - x^k\|^2,$$

где $f_k(x) = f(x^k) + \left(\nabla f(x^k), x - x^k\right) + \dfrac{\left(\nabla^2 f(x^k)(x - x^k), x - x^k\right)}{2}$.

Таким образом, метод: $x^{k+1} = x^k - \left(\nabla^2 f(x^k) + \alpha_k E\right)^{-1} \nabla f(x^k)$.

[3] При $\alpha_k = 0$ это будет метод Ньютона. При $\alpha_k \to \infty$ — направление движения стремится к антиградиенту. За счет выбора α_k можно добиться глобальной сходимости.

Алгоритм (начинаем с $\alpha_k \approx 10^4$):

1. Если $f(x^{k+1}) < f(x^k)$, то берем $\alpha_{k+1} < \alpha_k$ и $k = k + 1$.
2. Иначе $\alpha_k = \rho\alpha_k$, где $\rho > 1$, и пересчитываем этот шаг ещё раз.

Метод пригоден не только для выпуклых функций.

Многошаговая идея

Пытаемся учесть «предысторию» процесса; для ускорения сходимости. Методы

$$x^{k+1} = \varphi_k(x^k, \ldots, x^{k-s+1})$$

называются s-шаговыми.

Градиентный метод и метод Ньютона — одношаговые.

Метод тяжелого шарика

Двухшаговый метод:

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}),$$

где $\alpha > 0$, $\beta \geqslant 0$ — некоторые параметры.

Физическая аналогия: движение тела («тяжелого шарика») в потенциальном поле при наличии силы трения (или вязкости) описывается дифференциальным уравнением второго порядка $\mu \frac{d^2 x(t)}{dt^2} = -\nabla f\left(x(t)\right) - p\frac{dx(t)}{dt}$. Из-за потери энергии на трение тело в конце концов окажется в точке минимума потенциала $f(x)$.

Если рассмотреть разностный аналог уравнения, то получим итерационный метод.

Метод сопряженных градиентов [1, 2]

Параметры находятся из решения двумерной задачи оптимизации:

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k \left(x^k - x^{k-1}\right),$$

$$\{\alpha_k, \beta_k\} = \operatorname{argmin}_{\{\alpha_k, \beta_k\}} f\left(x^k - \alpha \nabla f(x^k) + \beta_k \left(x^k - x^{k-1}\right)\right).$$

Для случая квадратичной функции $f(x) = (Ax, x)/2 - (b, x)$, $A > 0$, есть явное решение:

$$\alpha_k = \frac{\left\|r^k\right\|^2 \left(Ap^k, p^k\right) - \left(r^k, p^k\right)\left(Ar^k, p^k\right)}{\left(Ar^k, r^k\right)\left(Ap^k, p^k\right) - \left(Ar^k, p^k\right)^2}, \quad r^k = \nabla f(x^k) = Ax^k - b,$$

$$\beta_k = \frac{\left\|r^k\right\|^2 \left(Ar^k, p^k\right) - \left(r^k, p^k\right)\left(Ar^k, p^k\right)}{\left(Ar^k, r^k\right)\left(Ap^k, p^k\right) - \left(Ar^k, p^k\right)^2}, \quad p^k = x^k - x^{k-1}.$$

Сходимость для квадратичной функции

Пусть начальное приближение x^0 произвольно, а x^1 получено из него методом скорейшего спуска:

$$x^1 = x^0 - \frac{\|r^0\|^2}{(Ar^0, r^0)}, \quad r^0 = \nabla f(x^0) = Ax^0 - b.$$

Theorem

Метод дает точку минимума квадратичной функции $f(x)$ рассматриваемого вида за число итераций, не превосходящее n.

Сопряженные градиенты: другая (вторая) форма записи

$$x^{k+1} = x^k + \alpha_k p^k,\ \alpha_k = \operatorname{argmin}_\alpha f\left(x^k + \alpha p^k\right),$$

$$p^k = -r^k + \beta_k p^{k-1},\ \beta_k = \frac{\left\|r^k\right\|^2}{\left\|r^{k-1}\right\|^2},$$

$$r^k = \nabla f(x^k),\ \beta_0 = 0$$

Theorem

Для случая квадратичной функции рассматриваемого вида при одинаковом x^0 оба метода определяют одну и ту же последовательность точек x^k.

Неквадратичные задачи: идея обновления

Для неквадратичных задач — несколько иная форма.
Вводится процедура обновления: время от времени шаг делается не по формуле, а как в начальной точке, т.е. по градиенту.
Например:

$$x^{k+1} = x^k + \alpha_k s^k, \ \alpha_k = \operatorname{argmin}_{\alpha \geqslant 0} f\left(x^k + \alpha s^k\right),$$

$$s^k = -r^k + \beta_k s^{k-1}, \ r^k = \nabla f(x^k),$$

$$\beta_k = \begin{cases} 0, \ k = 0, \ n, \ 2n, \ \ldots \\ \dfrac{\left\|r^k\right\|^2}{\left\|r^{k-1}\right\|^2}, \ k \neq 0, \ n, \ 2n, \ \ldots \end{cases}$$

Третья схема сопряженных градиентов

Для неквадратичных функций. Другое правило выбора β_k.

$$x^{k+1} = x^k + \alpha_k s^k,\ \alpha_k = \operatorname{argmin}_{\alpha \geqslant 0} f\left(x^k + \alpha s^k\right),$$

$$s^k = -r^k + \beta_k s^{k-1},\ \beta_k = \frac{\left(r^k,\, r^k - r^{k-1}\right)}{\left\|r^{k-1}\right\|^2},$$

$$r^k = \nabla f(x^k),\ \beta_0 = 0$$

- Здесь тоже возможны варианты с обновлением или без него.
- Для квадратичной функции последовательности x^k, порождаемые второй и третьей схемами совпадают.
- Считается, что для неквадратичного случая третья схема обычно дает более быструю сходимость.

Ускоренный метод Нестерова

Идея: инерция + использование антиградиента в прогнозируемой точке.
Итерационные формулы:

$$y_0 = x_0$$
$$x_1 = y_0 - \alpha \cdot f_x(y_0)$$
$$y_1 = x_1 + \beta(x_1 - x_0)$$
$$\dots$$
$$x_{k+1} = y_k - \alpha \cdot f_x(y_k)$$
$$y_{k+1} = x_{k+1} + \beta(x_{k+1} - x_k)$$

Сходимость и скорость сходимости

Пусть функция удовлетворяет условию Липшица на её градиент с константой L:

$$\|\nabla f(z_1) - \nabla f(z_2)\| \leqslant L\|z_1 - z_2\|.$$

Пусть

$$0 \leqslant \beta < 1, \quad 0 \leqslant \alpha < \frac{2(1-\beta)}{L}.$$

Тогда методы тяжелого шарика и Нестерова сходятся.

Верна оценка на число итераций N: $N = O\left(\frac{LR^2}{\sqrt{\varepsilon}}\right)$.

Стохастический градиентный спуск (см. книгу А.В.Гасникова в [3]

Идея: используем не $\nabla f(x)$, а некое, зависящее от случайной величины ξ, «приближение» — $g(x, \xi)$: $E_\xi g(x, \xi) = \nabla f(x)$.
Итерационная формула:

$$x_{k+1} = x_k - \alpha_k \cdot g(x_k, \xi_k).$$

Пример:

$$f(x) = \frac{1}{2m} \sum_{j=1}^{m} (x - y_i)^2$$

$$\Rightarrow \nabla f(x) = \frac{1}{m} \sum_{j=1}^{m} (x - y_i)$$

$$g(x, i) \equiv x - y_i.$$

Здесь, g — это в среднем градиент f.

Задание 10. Методы безусловной многомерной оптимизации

Список методов первого порядка (выбрать три):

- градиентный спуск
- наискорейший спуск
- метод тяжелого шарика
- метод сопряженных градиентов
- метод Нестерова
- стохастический градиентный спуск

Список методов второго порядка (выбрать один):

- метод Ньютона
- демпфированный метод Ньютона
- метод Левенберга–Марквардта

Для тестов: несколько примеров задач есть в [1] (глава 12).

Общая схема методов спуска

Задача $\min_{x \in \mathbb{R}^n} f(x)$.

Идея: строим последовательность точек $\{x^k\}$: $f(x^{k+1}) < f(x^k)$.
Различия:

- способы выбора направления убывания;
- способы выбора шага;

+ правила остановки процесса.

Ненулевой вектор $s \in \mathbb{R}^n$ называется направлением убывания функции $f(x)$ в точке $x \in \mathbb{R}^n$, если $f(x + \alpha s) < f(x)$ для достаточно малых $\alpha > 0$.
Множество всех направлений убывания функции $f(x)$ в точке x образуют конус — $\mathcal{K}_d(x)$.
Утверждение. Пусть функция $f(x)$ дифференцируема в точке $x \in \mathbb{R}^n$. Тогда:

1. для любого $s \in \mathcal{K}_d(x)$ выполнено $(\nabla f(x), s) \leqslant 0$;
2. если s удовлетворяет условию $(\nabla f(x), s) < 0$,то $s \in \mathcal{K}_d(x)$.

Далее рассматриваем итерационные методы спуска, где

$$x^{k+1} = x^k + \alpha_k s_k, \ s_k \in \mathcal{K}_d(x^k), \ \alpha_k > 0.$$

Если $\mathcal{K}_d(x^k) = \emptyset$, то процесс прерывается.

Правила выбора длины шага

Правило одномерной минимизации

В качестве α_k берется решение задачи

$$f(x^k + \alpha_k s_k) = \min_{\alpha \geqslant 0} f(x^k + \alpha s_k).$$

Если $\nabla f(x^{k+1}) \neq 0_n$, то геометрически решение этой задачи означает, что x^{k+1} является точкой касания луча, задаваемого направлением s_k, с поверхностью уровня функции $f(x)$, проходящей через точку x^{k+1}.

В общем случае задачу минимизации решить непросто, поэтому её решают приближенно: вместо поиска минимума на луче ищут минимум на отрезке $[0, \bar{\alpha}]$.

Правило Армихо

Это — приближенный способ нахождения шага α_k.
Пусть $f(x)$ дифференцируема в точке x^k. Задаются два числа:
$0 < \varepsilon < 1$ и $0 < \theta < 1$ и выбирают начальное значение длины шага $\bar{\alpha}$.
Полагают $\alpha = \bar{\alpha}$. Выбор α_k проводится двухэтапной процедурой:

1. проверка выполнения условия (неравенство Армихо)

$$f(x^k + \alpha s_k) - f(x^k) \leqslant \varepsilon\alpha \left(\nabla f(x^k),\, s_k\right);$$

2. если неравенство не выполняется, то заменяем α на $\alpha := \theta\alpha$ и повторяем первый этап; иначе заканчиваем процесс и полагаем $\alpha_k = \alpha$.

Правило Голдстейна

Задаются два параметра $0 < \varepsilon_1 < 1$ и $0 < \varepsilon_2 < 1$, причем $\varepsilon_1 < \varepsilon_2$.
Шаг α на k-й итерации подбирается таким образом, чтобы было верно неравенство

$$\varepsilon_1 \leqslant \frac{f(x^k + \alpha s_k) - f(x^k)}{\alpha \left(\nabla f(x^k),\, s_k\right)} \leqslant \varepsilon_2.$$

Левое неравенство — это идейно правило Армихо.
Правое — чтобы шаг не был достаточно малым.

Правило априорного выбора

Последовательность шагов $\{\alpha_k\}$ задается такая, чтобы

$$\alpha_k > 0, \sum_{k=0}^{\infty} \alpha_k = \infty, \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

На некоторых шагах метод может перестать быть методом спуска.
Используется при минимизации негладких функций.

Где что:

В [1] методы для недифференцируемых функций (глава 5):

- субградиентный метод, §3;
- многошаговые методы, §4.

Можно почитать и про влияние помех.

В [2] многоэкстремальные методы (глава 11):

- метод неравномерных покрытий;
- метод секущих углов.

Метод сопряженных градиентов, возможно, описан человеколюбивее.

В [4]:

- метод Пауэлла (безградиентный);
- динамический метод (сравните с методом тяжелого шарика :)).

Есть полезные практические указания.