УДК 517.9

Градиентные методы минимизации [1]

М. В. Балашов (ИПУ РАН, Москва, Россия)
balashov73@mail.ru

В работе обсуждается современное состояние дел в градиентных методах решения задач поиска минимума функции на множестве: методе проекции градиента и методе условного градиента. Будут сформулированы результаты о линейной сходимости для выпуклых задач, а также рассмотрены возможные обобщения этих результатов на невыпуклый случай.

Ключевые слова: Метод проекции градиента, метод условного градиента, сильная выпуклость, липшицев градиент, метрическая регулярность, негладкий анализ.

Gradient minimization methods [1]

M. V. Balashov (ICS RAS, Moscow, Russia)
balashov73@mail.ru

In the paper, we discuss the current state of arts in gradient methods for solving problems of finding the minimum of a function on a set: about the gradient projection method and about the conditional gradient method. Results on linear convergence for convex problems will be formulated, and possible generalizations of these results to the non-convex case will be considered.

Keywords: gradient projection method, conditional gradient method, strong convexity, Lipschitz continuous gradient, metric regularity, nonsmooth analysis.

Введение

В $\mathbb{R}^n$ через (x, y) обозначим скалярное произведение векторов $x, y \in \mathbb{R}^n$ и соответствующую евклидову норму $\|x\| = \sqrt{(x, x)}$. Для функции $f : \mathbb{R}^n \to \mathbb{R}$ через $f'(x)$ будем обозначать градиент Фреше f в точке x.

Рассмотрим задачу

$$f_0 = \min_{x \in \mathcal{Q}} f(x), \tag{1}$$

где $\mathcal{Q} \subset \mathbb{R}^n$ замкнутое подмножество и $f : \mathbb{R}^n \to \mathbb{R}$ — функция с непрерывным по Липшицу градиентом. Через $\Omega_0 \subset \mathcal{Q}$ будем обозначать непустое множество решений (1). Условие Липшица для градиента $f'(x)$ достаточно требовать в некоторой равномерной окрестности множества $\mathcal{Q}$. Везде ниже мы будем предполагать, что решение по x задачи (1) существует, т.е. $\Omega_0 \neq \emptyset$ и $f_0 \in \mathbb{R}$. Пусть $P_{\mathcal{Q}} x$ — метрическая проекция точки x на множество $\mathcal{Q}$ и $\varrho(x, \mathcal{Q})$ — расстояние от точки x до множества $\mathcal{Q}$.

В докладе будут обсуждаться условия линейной сходимости, т.е. сходимости со скоростью геометрической прогрессии со знаменателем < 1, метода проекции градиента

$$x_1 \in \mathcal{Q}, \qquad x_{k+1} \in P_{\mathcal{Q}}\left(x_k - tf'(x_k)\right), \quad t > 0, \tag{2}$$

и метода условного градиента $x_1 \in \mathcal{Q}$

$$z_k \in \operatorname{Arg}\max_{x\in\mathcal{Q}}(-f'(x_k), x), \qquad x_{k+1} = (1-t_k)x_k + t_k z_k, \quad t_k \in [0,1]. \tag{3}$$

В частности, точку x_{k+1} по точкам x_k, z_k можно выбирать в алгоритме (3) по правилу

$$x_{k+1} \in \operatorname{Arg}\min_{x\in[x_k,z_k]} f(x). \tag{4}$$

Для простоты мы будем рассматривать ситуацию в $\mathbb{R}^n$, хотя большинство фактов легко переносится на случай вещественного гильбертова пространства.

Общая ситуация. Монотонность алгоритмов. Выпуклый случай

Начнём с рассмотрения алгоритмов (2) и (3) для произвольного замкнутого множества $\mathcal{Q}$ и достаточно произвольного вида функции f.

Напомним, что функция $f : U \to \mathbb{R}$ слабо вогнута на выпуклом подмножестве $U \subset \mathbb{R}^n$ с константой $L_1 > 0$, если функция $h(x) = f(x) - \frac{L_1}{2}\|x\|^2$ вогнута (выпукла вверх) на U. Дифференцируемость функции f не предполагается. *Супердифференциал Фреше* функции f в точке $x_0 \in U$ есть

$$\partial_F^+ f(x_0) = \left\{ p \in \mathbb{R}^n : \quad \limsup_{x\to x_0} \frac{f(x) - f(x_0) - (p, x - x_0)}{\|x - x_0\|} \leq 0 \right\}.$$

Для слабо вогнутой с константой L_1 функции f имеет место равенство [1, Proposition 1.107(i)]

$$\partial_F^+ f(x_0) = \partial^+ h(x_0) + L_1 x,$$

где $\partial^+ h(x_0)$ супердифференциал вогнутой функции h. Будем говорить, что функция удовлетворяет условию верхней квадратичной аппроксимации с константой L_1, если

$$f(x) \leq f(x_0) + (f'(x_0), x - x_0) + \frac{L_1}{2}\|x - x_0\|^2, \qquad \forall x_0, x, \ \forall f'(x_0) \in \partial_F^+ f(x_0).$$

Из определения слабой вогнутости легко вытекает, что слабо вогнутая функция (и только слабо вогнутая функция) удовлетворяет условию верхней квадратичной аппроксимации с одной константой L_1.

Лемма 1. *Пусть функция $f : \mathbb{R}^n \to \mathbb{R}$ слабо вогнутая с константой $L_1 > 0$, множество $\mathcal{Q}$ замкнуто, $x_1 \in \mathcal{Q}$, $t \in \big(0, \frac{1}{L_1}\big)$, суперградиент Фреше $f'(x_1) \in \partial_F^+ f(x_1)$ и $x_2 \in P_{\mathcal{Q}}\big(x_1 - tf'(x_1)\big)$ произвольны. Тогда*

$$f(x_1) - f(x_2) \geq \frac{1}{2}\left(\frac{1}{t} - L_1\right)\|x_2 - x_1\|^2. \tag{5}$$

Заметим, что в лемме 1 множество $\mathcal{Q}$, как и функция f, не выпукло, поэтому проекция $P_{\mathcal{Q}}\big(x_1 - tf'(x_1)\big)$ вообще говоря не одноточечна.

Доказательство. Зафиксируем произвольным образом суперградиент Фреше $f'(x_1) \in \partial_F^+ f(x_1)$. Рассмотрим квадратичную функцию

$$\psi(x) = f(x_1) + \big(f'(x_1), x - x_1\big) + \frac{1}{2t}\|x - x_1\|^2.$$

В силу условия слабой вогнутости f с константой L_1 выполнена квадратичная аппроксимация f, откуда для всякого $x \in \mathbb{R}^n$ выполнена оценка

$$\psi(x) \geq f(x) + \frac{1}{2}\left(\frac{1}{t} - L_1\right)\|x - x_1\|^2.$$

Для функции $\psi(x)$ точка $x_1 - tf'(x_1)$ является глобальным минимумом. Из того, что линии уровня $\psi(x) = C$ — сферы с центром $x_1 - tf'(x_1)$, получаем $\psi(x_2) \leq \psi(x_1)$. Отсюда

$$f(x_1) = \psi(x_1) \geq \psi(x_2) \geq f(x_2) + \frac{1}{2}\Big(\frac{1}{t} - L_1\Big)\|x_2 - x_1\|^2.$$

$\square$

Аналогично, из условия верхней квадратичной аппроксимации в методе (3), для всякого для $\varepsilon \in (0, 1)$ и $t \in (\varepsilon, (2-\varepsilon)/L_1)$ найдётся константа $C(\varepsilon) > 0$ такая, что [2, §6]

$$f(x_{k+1}) - f(x_k) \leq -C(\varepsilon) \min\left\{ \big(f'(x_k), x_k - z_k\big), \frac{\big(f'(x_k), x_k - z_k\big)^2}{\|x_k - z_k\|^2} \right\} \leq 0.$$

Таким образом, для слабо вогнутой функции алгоритм (3) также является строго монотонным.

В дальнейшем мы будем считать, что функция f имеет липшицев градиент с константой L_1. Известно, что такие функции слабо вогнутые с той же константой L_1 [3, теорема 2.1.2]. Приведённые результаты гарантирует монотонность алгоритма для такой функции.

При условии липшицевой дифференцируемости и выпуклости (выпуклости вниз) f, а также выпуклости и компактности множества $\mathcal{Q}$, оценка в алгоритмах (2) и (3) по функции имеет вид

$$f(x_k) - f_0 \leq \frac{C}{k}$$

для некоторой константы $C > 0$, зависящей от f и $\mathcal{Q}$ [2]. Пример функций $f_m(x) = x^{2m}$, $m \in \mathbb{N}$, в задаче $\min\limits_{x \in \mathbb{R}^n} f_m(x)$ (с очевидным решением $x_0 = 0$) показывает, что скорость сходимости $\|x_k - x_0\|$ в методе градиентного спуска может быть сколь угодно медленной. Аналогичный пример можно привести для алгоритма (3).

Градиентный спуск в случае $\mathcal{Q} = \mathbb{R}^n$

Будем говорить, что функция $f\colon \mathbb{R}^n \to \mathbb{R}$ удовлетворяет *условию Лежанского-Поляка-Лоясевича (условию LPL)* [4], если f дифференцируема по Фреше, множество $\Omega_0 = \operatorname{Arg}\min_{x \in \mathbb{R}^n} f(x)$ ее глобальных минимумов непусто и существует такая константа $\mu > 0$, что для всех $x \in \mathbb{R}^n$

$$\|f'(x)\|^2 \geq \mu\big(f(x) - f(x_0)\big) \qquad \forall x_0 \in \Omega_0. \tag{6}$$

Совместно с формулой (5) леммы 1 для липшицево дифференцируемой функции f, удовлетворяющей условию (6), получаем для шага $0 < t < 1/L_2$

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2}\left(\frac{1}{t} - L_1\right)\|tf'(x_k)\|^2 \geq \frac{\mu}{2}\left(t - t^2 L_1\right)\left(f(x_k) - f_0\right),$$

и для $q = q(\mu, t) = \frac{\mu}{2}\left(t - t^2 L_1\right)$ и $\varphi_k = f(x_k) - f_0$ имеем

$$\varphi_k - \varphi_{k+1} \geq q\varphi_k, \quad \varphi_{k+1} \leq (1-q)\varphi_k,$$

т.е. получаем линейную сходимость $\{\varphi_k\}$ к нулю. По точке

$$\|x_{k+1} - x_k\|^2 = t^2\|f'(x_k)\|^2 \leq \frac{\varphi_k}{\frac{1}{2}(\frac{1}{t} - L_1)}$$

также получаем линейную сходимость.

Отметим, что сильно выпуклая с константой $\varkappa > 0$ функция (т.е. такая функция f, что функция $f(x)-\frac{\varkappa}{2}\|x\|^2$ выпуклая вниз) удовлетворяет условию в духе (LPL) вида

$$\|p\|^2 \geq \varkappa\big(f(x) - f_0\big) \quad \forall p \in \partial f(x),$$

где $\partial f(x) = \{p \in \mathbb{R}^n \ : \ f(z) \geq f(x) + (p, z-x), \quad \forall z \in \mathbb{R}^n\}$ есть субдифференциал функции f в точке x (дифференцируемость f не предполагается). Существуют также невыпуклые функции, удовлетворяющие условию (LPL). В градиентном спуске можно эффективно применять шаг Армихо, для вычисления которого не нужно знание констант.

Минимизация на многообразиях

Рассмотрим частный случай задачи (1), когда множество $\mathcal{Q}$ является гладким компактным многообразием без края, а функция f является вещественно-аналитической, т.е. для каждой точки найдётся окрестность, где функция представима некоторым сходящимся степенным рядом. Тогда для всякой точки $x \in \mathcal{Q}$ найдётся $\delta > 0$, $\alpha \in (1, 2]$ и $\mu > 0$ такие, что для всякого $y \in \mathcal{Q}$, $\|y-x\| \leq \delta$, выполнено неравенство в духе неравенства Лоясевича [5, Proposition 2.2]

$$\mu|f(y) - f(x)| \leq \|P_{T_y} f'(y)\|^\alpha, \tag{7}$$

где T_y — касательное подпространство в точке $y \in \mathcal{Q}$ к многообразию $\mathcal{Q}$.

Компактное гладкое многообразие без края является проксимально гладким с некоторой константой $R > 0$ [3, теорема 1.19.1]. Последнее эквивалентно тому, что для всякого $r \in (0, R)$ и точек $x_0, x_1 \in \mathbb{R}^n$, $0 < \varrho(x_i, \mathcal{Q}) < r$, $i = 0, 1$, выполнено условие Липшица $\|P_{\mathcal{Q}} x_0 - P_{\mathcal{Q}} x_1\| \leq \frac{R}{R-r}\|x_0 - x_1\|$ [6]. Напомним, что основные матричные многообразия: Штифеля (матрицы $X \in \mathbb{R}^{n\times k}$: $X^T X = I_{k\times k}$), Грассмана (может быть реализовано как множество проекторов XX^T, где $X \in \mathbb{R}^{n\times k}$ — матрица многообразия Штифеля) и т.д. проксимально гладкие; многообразие Штифеля с константой $R = 1$, а многообразие Грассмана с $R = 1/\sqrt{2}$ для всех n, k, $k \leq n$. Существуют эффективные формулы для проектирования точки (матрицы) на матричные многообразия через сингулярное разложение матриц [7, 8]. Можно также доказать, что аналогом формулы (5) для двух последовательных шагов метода проекции градиента (2), который принимает вид $x_{k+1} = P_S(x_k - tP_{T_{x_k}} f'(x_k))$, на гладком и проксимально гладком многообразии является неравенство [9, Theorem 2]

$$f(x_{k+1}) - f(x_k) \leq -\|P_{T_{x_k}} f'(x_k)\|^2 \left(t - t^2\left(\frac{L}{R} + \frac{L_1}{2}\right)\right), \tag{8}$$

где L и L_1 — константы Липшица f и f' соответственно, а R — константа проксимальной гладкости S. Формулы (7) и (8) аналогично случаю безусловной минимизации дают линейную сходимость указанного алгоритма к локальному минимуму с сублинейной скоростью при $\alpha < 2$ и с линейной скоростью при $\alpha = 2$ [10]. Возможно эффективно применять шаг Армихо [8].

Общая ситуация

Нестеров и др. доказали [11], что если в задаче (1) с выпуклой и липшицево дифференцируемой функцией f и выпуклым компактным множеством $\mathcal{Q}$ выполнено условие квадратичного роста

$$\exists \nu > 0 \qquad f(x) - f_0 \geq \nu \varrho^2(x, \Omega_0) \quad \forall x \in \mathcal{Q}, \tag{9}$$

то алгоритм (2) с достаточно малым фиксированным шагом $t > 0$ сходится к решению (1) с линейной скоростью. Про алгоритм (3) давно известно [2], что для его линейной сходимости достаточно неравенство $\inf\limits_{x \in \mathcal{Q}} \|f'(x)\| > 0$ и условие сильной выпуклости множества $\mathcal{Q}$ с радиусом $R > 0$, т.е. $\mathcal{Q} = \bigcap\limits_{x \in X} B_R(x)$ есть пересечение замкнутых евклидовых шаров $B_R(x)$ для некоторого $R > 0$ и произвольного множества центров шаров $X \subset \mathbb{R}^n$. В работе [12] было отмечено, что для линейной сходимости достаточно, чтобы условие сильной выпуклости выполнялось локально, в точке-решении $x_0 \in \mathcal{Q}$ множество $\mathcal{Q}$ должно иметь локальный модуль выпуклости второго порядка.

Кроме того, известно, что если в задаче (1) функция (м.б. невыпуклая) имеет липшицев градиент с константой $L_1 > 0$, множество $\mathcal{Q}$ сильно выпукло с радиусом $R > 0$ и $\|f'(x)\| \geq m > 0$ при всех $x \in \mathcal{Q}$, то при условии $\frac{RL_1}{m} < 1$ алгоритм (3) сходится при выборе $x_{k+1} = z_k$. Этот результат и некоторые его уточнения могут быть найдены в [4]. Мы также рекомендуем недавний обзор [13], посвященный методу условного градиента.

При попытке отказа от выпуклости мы встречаемся с определенными трудностями. Так, условие (9) не является в общем случае достаточным для сходимости алгоритма (2) в невыпуклом случае, т.к. могут существовать локальные экстремумы. Про алгоритм (3) автору не известны какие-либо условия его сходимости с невыпуклым множеством $\mathcal{Q}$.

Для решения ряда задач вида (1) в выпуклом и невыпуклом случае мы планируем обсудить опорное условие сильной выпуклости для компактного и в общем случае невыпуклого множества $\mathcal{Q}$ [14]. Опорное условие для компактного множества $\mathcal{Q}$ определяется направлением

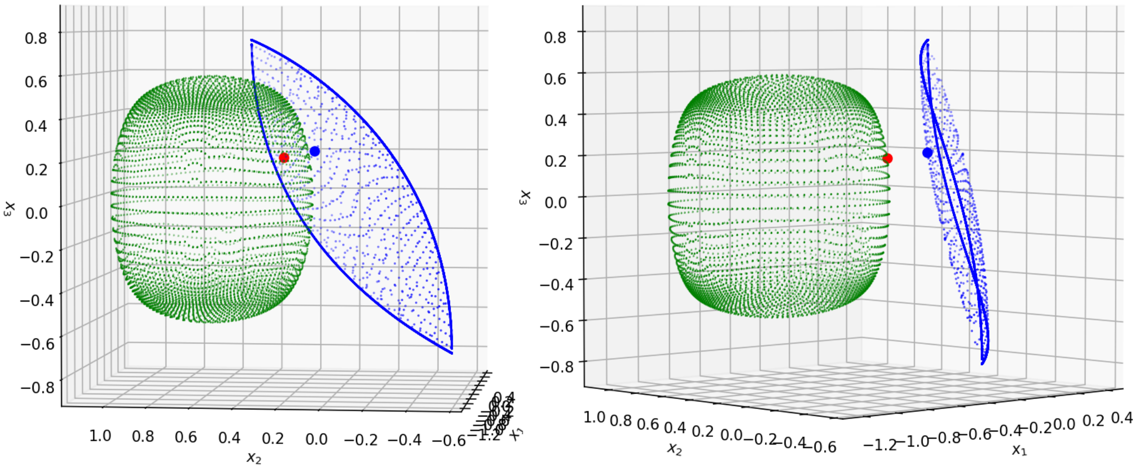

Рис. 1: Ближайшие точки множеств $\mathcal{R}(3)$ (множество справа) и $\mathcal{M}$ (множество слева), разный ракурс. Решается задача $\min\limits_{\|p\|=1} s(p, \mathcal{R}(3) + (-\mathcal{M}))$.

единичного вектора $p \in \mathbb{R}^n$ и радиусом R и означает, что

$$\mathcal{Q} \subset B_R(\mathcal{Q}(p) - Rp), \qquad \mathcal{Q}(p) = \arg\max_{x \in \mathcal{Q}} (p, x). \tag{10}$$

Оказывается, что это условие и некоторое соотношение констант достаточно для сходимости ряда алгоритмов вида (2) и (3) с $x_{k+1} = z_k$.

Приведём пример, см. рис. 1. Рассмотрим в $\mathbb{R}^3$ множество достижимости $\mathcal{R}(t) = \int\limits_0^t e^{As} U \, ds$ для $t = 3$, $U = [-e_3, e_3]$. Матрица A имеет вид жордановой клетки 3×3 с собственным значением $\lambda = -1.3$, а $\mathcal{M}$ есть сумма Минковского некоторых двух эллипсоидов. Ниже в таблице 1 приведена скорость сходимости опорной функции $f(p) = s(p, \mathcal{R}(3) + (-\mathcal{M}))$ к минимуму (т.е. $f(p_k) - f_0 \le Cq^k$) в алгоритме (2) при $\mathcal{Q} = \{p \in \mathbb{R}^n \ : \ \|p\| = 1\}$ для разных шагов t. При $t > 0.4$ алгоритм может разойтись.

Таблица 1: Скорость линейной сходимости

Шаг t	0.05	0.1	0.15	0.2	0.25	0.3	0.35
Показатель q	0.914	0.829	0.744	0.660	0.574	0.633	0.862

СПИСОК ЛИТЕРАТУРЫ

[1] *Mordukhovich B.S.* Variational Analysis and Generalized Differentiation I: Basic Theory. Berlin: Springer, 2005.

[2] *Левитин Е.С., Поляк Б.Т.* Методы минимизации при наличии ограничений // Ж. выч. матем. и матем. ф. 1966. Т. 6. № 5. С. 787–823.

[3] *Иванов Г.Е.* Слабо выпуклые функции и множества. М.: Физматлит, 2006.

[4] *Balashov M. V., Polyak B. T., Tremba A. A.* Gradient Projection and Conditional Gradient Methods for Constrained Nonconvex Minimization // Numerical Functional Analysis and Optimization. 2020 V. 41. №7. P. 822–849.

[5] *Schneider R., Uschmaev A.* Convergence results for projected line-search methods on varieties of low-rank matricies via Lojasiewicz inequality // SIAM J. Opt. 2015. V. 25. № 1. P. 622–646.

[6] *Vial J.-P.* Strong and weak convexity of sets and functions // Mathematics of Operations Research. 1983. Vol. 8. No. 2. P. 231–259.

[7] *Балашов М.В.* Метод проекции градиента на матричных многообразиях // Ж. выч. матем. и матем. ф. 2020. Т. 60. № 9. С. 1453–1461.

[8] *Балашов М.В., Камалов Р. А.* Метод проекции градиента с шагом Армихо на многообразиях // Ж. выч. матем. и матем. ф. 2021. Т. 61. № 11. С. 1814–1824.

[9] *Balashov M. V.* The gradient projection algorithm for smooth sets and functions in nonconvex case // Set-Valued Var. Anal. 2021. V. 29. P. 341–360.

[10] *Balashov M. V.* The Lezanski – Polyak – Lojasiewicz inequality and the convergence of the gradient projection algorithm // Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика. 2023. Т. 23. Вып. 1. С. 4–10.

[11] *Necoara I., Nesterov Yu., Glineur F.* Linear convergence of first order methods for non-strongly convex optimization // Math. Prog. Series A. 2019. Vol. 175. No. 1–2. P. 69–107.

[12] *Veliov V.M.* On the convexity of integrals of multivalued mappings: application in control theory // J. of Opt. Theory and Appl. 1987. Vol. 54. No. 3. P. 541–563.

[13] *Braun G., Carderera A., Combettes C.W., Hassani H., Karbasi A., Mokhtari A., Pokutta S.* Conditional gradient methods // arXiv:2211.14103v1. 2022.

[14] *Балашов М.В.* Достаточные условия линейной сходимости одного алгоритма для нахождения метрической проекции точки на выпуклый компакт // Матем. зам. 2023. Т. 113. № 5. С. 655–666.