**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2024 Т. 16 № 7 С. 1829–1840**
DOI: 10.20537/2076-7633-2024-16-7-1829-1840

Ки&М

СПЕЦИАЛЬНЫЙ ВЫПУСК

УДК: 519.85

Регуляризация и ускорение метода Гаусса – Ньютона

Н. Е. Юдин[1,2,3,4,5,6,a], А. В. Гасников[1,2,b]

[1]Университет Иннополис,
Россия, 420500, г. Иннополис, ул. Университетская, д. 1
[2]Национальный исследовательский университет «Московский физико-технический институт»,
Россия, 141701, Московская область, г. Долгопрудный, Институтский пер., д. 9
[3]Национальный исследовательский университет «Высшая школа экономики»,
Россия, 109028, г. Москва, Покровский бульвар, д. 11
[4]Федеральное государственное бюджетное учреждение науки «Институт проблем передачи информации им. А. А. Харкевича» Российской академии наук,
Россия, 127051, г. Москва, Большой Каретный пер., д. 19, стр. 1
[5]Федеральное государственное бюджетное учреждение науки «Институт системного программирования им. В. П. Иванникова» Российской академии наук,
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25
[6]Федеральный исследовательский центр «Информатика и управление» Российской академии наук,
Россия, 119333, г. Москва, ул. Вавилова, д. 44, корп. 2

E-mail: [a] iudin.ne@phystech.edu, [b] gasnikov@yandex.ru

Получено 28.10.2024, после доработки — 12.11.2024
Принято к публикации 25.11.2024

Предлагается семейство методов Гаусса – Ньютона для решения оптимизационных задач и систем нелинейных уравнений, основанное на идеях использования верхней оценки нормы невязки системы уравнений и квадратичной регуляризации. В работе представлено развитие схемы метода трех квадратов с добавлением моментного члена к правилу обновления искомых параметров в решаемой задаче. Получившаяся схема обладает несколькими замечательными свойствами. Во-первых, в работе алгоритмически описано целое параметрическое семейство методов, минимизирующих функционалы специального вида: композиции невязки нелинейного уравнения и унимодального функционала. Такой функционал, целиком согласующийся с парадигмой «серого ящика» в описании задачи, объединяет в себе большое количество решаемых задач, связанных с приложениями в машинном обучении, с задачами восстановления регрессионной зависимости. Во-вторых, полученное семейство методов описывается как обобщение нескольких форм алгоритма Левенберга – Марквардта, допускающих реализацию в том числе и в неевклидовых пространствах. В алгоритме, описывающем параметрическое семейство методов Гаусса – Ньютона, используется итеративная процедура, осуществляющая неточное параметризованное проксимальное отображение и сдвиг с помощью моментного члена. Работа содержит детальный анализ эффективности предложенного семейства методов Гаусса – Ньютона, выведенные оценки учитывают количество внешних итераций алгоритма решения основной задачи, точность и вычислительную сложность представления локальной модели и вычисления оракула. Для семейства методов выведены условия сублинейной и линейной сходимости, основанные на неравенстве Поляка – Лоясиевича. В обоих наблюдаемых режимах сходимости локально предполагается наличие свойства Липшица у невязки нелинейной системы уравнений. Кроме теоретического анализа схемы, в работе изучаются вопросы ее практической реализации. В частности, в проведенных экспериментах для субоптимального шага приводятся схемы эффективного вычисления аппроксимации наилучшего шага, что позволяет на практике улучшить сходимость метода по сравнению с оригинальным методом трех квадратов. Предложенная схема объединяет в себе несколько существующих и часто используемых на практике модификаций метода Гаусса – Ньютона, в добавок к этому в работе предложена монотонная моментная модификация семейства разработанных методов, не замедляющая поиск решения в худшем случае и демонстрирующая на практике улучшение сходимости метода.

Ключевые слова: системы нелинейных уравнений, невыпуклая оптимизация, метод Гаусса – Ньютона, условие Поляка – Лоясиевича, оценка сложности

Исследование выполнено за счет гранта Российского научного фонда (проект № 21-71-30005, https://rscf.ru/project/21-71-30005/).

COMPUTER RESEARCH AND MODELING
2024 VOL. 16 NO. 7 P. 1829–1840
DOI: 10.20537/2076-7633-2024-16-7-1829-1840

Ки&М

SPECIAL ISSUE

UDC: 519.85

Regularization and acceleration of Gauss – Newton method

N. E. Yudin[1,2,3,4,5,6,a], A. V. Gasnikov[1,2,b]

[1]Innopolis University,
1 Universitetskaya st., Innopolis, 420500, Russia
[2]National Research University "Moscow Institute of Physics and Technology",
9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia
[3]HSE University,
11 Pokrovsky bul., Moscow, 109028, Russia
[4]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute),
19/1 Bolshoy Karetny per., Moscow, 127051, Russia
[5]Ivannikov Institute for System Programming of the Russian Academy of Sciences,
25 A. Solzhenitsyn st., Moscow, 109004, Russia
[6]Federal Research Center "Informatics and Control" of the Russian Academy of Sciences,
44/2 Vavilova st., Moscow, 119333, Russia

E-mail: [a] iudin.ne@phystech.edu, [b] gasnikov@yandex.ru

Received 28.10.2024, after completion — 12.11.2024
Accepted for publication 25.11.2024

We propose a family of Gauss – Newton methods for solving optimization problems and systems of nonlinear equations based on the ideas of using the upper estimate of the norm of the residual of the system of nonlinear equations and quadratic regularization. The paper presents a development of the «Three Squares Method» scheme with the addition of a momentum term to the update rule of the sought parameters in the problem to be solved. The resulting scheme has several remarkable properties. First, the paper algorithmically describes a whole parametric family of methods that minimize functionals of a special kind: compositions of the residual of a nonlinear equation and an unimodal functional. Such a functional, entirely consistent with the «gray box» paradigm in the problem description, combines a large number of solvable problems related to applications in machine learning, with the regression problems. Secondly, the obtained family of methods is described as a generalization of several forms of the Levenberg – Marquardt algorithm, allowing implementation in non-Euclidean spaces as well. The algorithm describing the parametric family of Gauss – Newton methods uses an iterative procedure that performs an inexact parametrized proximal mapping and shift using a momentum term. The paper contains a detailed analysis of the efficiency of the proposed family of Gauss – Newton methods; the derived estimates take into account the number of external iterations of the algorithm for solving the main problem, the accuracy and computational complexity of the local model representation and oracle computation. Sublinear and linear convergence conditions based on the Polak – Lojasiewicz inequality are derived for the family of methods. In both observed convergence regimes, the Lipschitz property of the residual of the nonlinear system of equations is locally assumed. In addition to the theoretical analysis of the scheme, the paper studies the issues of its practical implementation. In particular, in the experiments conducted for the suboptimal step, the schemes of effective calculation of the approximation of the best step are given, which makes it possible to improve the convergence of the method in practice in comparison with the original «Three Square Method». The proposed scheme combines several existing and frequently used in practice modifications of the Gauss – Newton method, in addition, the paper proposes a monotone momentum modification of the family of developed methods, which does not slow down the search for a solution in the worst case and demonstrates in practice an improvement in the convergence of the method.

Keywords: systems of nonlinear equations, non-convex optimization, Gauss – Newton method, Polyak – Lojasiewicz condition, complexity estimate

Citation: *Computer Research and Modeling*, 2024, vol. 16, no. 7, pp. 1829–1840 (Russian).

The research was supported by Russian Science Foundation (project No. 21-71-30005, https://rscf.ru/en/project/21-71-30005/).

Введение

Рассмотрим задачу решения гладкой системы нелинейных уравнений $F\colon \mathbb{R}^n \to \mathbb{R}^m$:

$$F(x) = \mathbf{0}_m, \quad \mathbf{0}_m^\top = (0, \ldots, 0). \tag{1}$$

Данная задача является фундаментальной в численных методах и в методах оптимизации [Голиков, Евтушенко, Капорин, 2019; Nesterov, 2018; Гасников, 2020]. Решение системы рассматривается в ключе задачи безусловной оптимизации невязки на основе евклидовой нормы:

$$\min_{x\in\mathbb{R}^n}\left\{f_1(x) \overset{\text{def}}{=} \|F(x)\|\right\}. \tag{2}$$

Задача (2) решается с помощью применения *метода Гаусса – Ньютона*. Одно из неоспоримых преимуществ метода выражено в свойстве иметь линейную или даже суперлинейную скорость сходимости, с использованием только информации о первых производных [Cai et al., 2019]. Это нередко позволяет избежать части негативных последствий, таких как увеличение необходимого количества итераций для достижения решения требуемой точности. Увеличение сложности решения относительно итераций на практике возможно посредством прямой минимизации

$$f_2(x) \overset{\text{def}}{=} (f_1(x))^2,$$

в том числе при использовании *методов доверительной области* или *квазиньютоновских методов* для минимизации квадрата оптимизируемой функции в (2) с применением различных стратегий [Le et al., 2020; Botev, Ritter, Barber, 2017; Ren, Goldfarb, 2019; Thiele, Araya-Polo, Hohl, 2020]. Более того, возведение в квадрат оптимизируемой функции часто приводит к росту вычислительной нестабильности методов оптимизации [Nesterov, 2007]. Сам по себе метод Гаусса – Ньютона примечателен тем, что на практике он может работать как метод второго порядка, формально являясь методом первого порядка, то есть методом, использующим информацию о первых производных вектор-функции F. Данная работа одновременно на теоретическом и на практическом уровне развивает метод Гаусса – Ньютона. Этот метод чрезвычайно полезен для задач умеренных размерностей, когда количество оптимизируемых переменных исчисляется несколькими сотнями. Несмотря на полезность указанного метода, в литературе мало анализа посвящено практичным вариантам самого метода, особенно на русском языке. Таким образом, текущая работа является развитием слабо изученного в теории и важного на практике класса методов, одной из немногих работ с тонким анализом модификаций на стыке теории и практики оптимизации невыпуклых задач, особенно с использованием приемов выпуклой релаксации.

В работе представлена модификация метода Гаусса – Ньютона с сильно выпуклой параметризованной мажорантой нормы невязки системы нелинейных уравнений. Прообраз метода под названием «метод трех квадратов» предложил Юрий Евгеньевич Нестеров в [Nesterov, 2020]. С использованием предложенной мажоранты построены алгоритмы детерминированной оптимизации для решения задачи (2) с практической возможностью получить ускорение за счет введения моментного члена без ущерба для скорости сходимости в худшем случае. Для представленного алгоритма решения задачи (2) дан анализ сходимости с неасимптотическими оценками относительно потенциала уровня $\epsilon > 0$, используемого в качестве индикатора сходимости итеративного процесса.

В втором разделе представлены необходимые для изложения результатов работы условные обозначения. В третьем разделе представлен анализ метода. Вводятся необходимые предположения, определяется форма мажоранты оптимизируемой функции, на основе которой строится весь метод, предлагается сама схема метода 1 с моментным ускорением. Для метода доказываются оценки глобальной сублинейной и линейной сходимости. В четвертом разделе представлены численные результаты для различных модельных задач, демонстрируется, как теоретически обоснованные модификации хорошо работают на практике. В приложении приведены доказательства полученных в данной работе утверждений, графики с результатами проведенных экспериментов и таблицы.

Условные обозначения

Для гладкой по x функции $f\colon \mathbb{R}^n \to \mathbb{R}^m$ обозначим вычисленную в точке $x \in \mathbb{R}^n$ первую производную по x как $\nabla_x f(x)$ (в случае отсутствия неоднозначности при определении переменной дифференцирования индексация у ∇ опускается). С помощью стандартного скалярного произведения определим сопряженный оператор $A^\top\colon \mathbb{R}^m \to \mathbb{R}^n$ для оператора $A\colon \mathbb{R}^n \to \mathbb{R}^m$:

$$\langle u, Ax\rangle = \left\langle A^\top u, x\right\rangle, \quad u \in \mathbb{R}^m,\ x \in \mathbb{R}^n.$$

Определим операторную норму для линейного оператора $A\colon \mathbb{R}^n \to \mathbb{R}^m$ как максимальное сингулярное число матрицы оператора $\sigma_{\max}(A)$:

$$\|A\| \overset{\text{def}}{=} \sigma_{\max}(A) = \max_{x\in\mathbb{R}^n}\{\|Ax\|\colon \|x\| \leqslant 1\} = \sqrt{\lambda_{\max}\left(AA^\top\right)} = \sqrt{\lambda_{\max}\left(A^\top A\right)},$$

где $\lambda_{\max}(\cdot)$ — максимальное собственное значение оператора. Также введем минимальное сингулярное число матрицы данного оператора A:

$$\sigma_{\min}(A) \overset{\text{def}}{=} \min_{x\in\mathbb{R}^n}\{\|Ax\|\colon \|x\| = 1\} = \sqrt{\lambda_{\min}\left(A^\top A\right)},$$

где $\lambda_{\min}(\cdot)$ — минимальное собственное значение оператора. Для векторнозначного отображения $F\colon \mathbb{R}^n \to \mathbb{R}^m$ обозначим матрицу Якоби $F'(x)$, вычисленную в точке $x \in \mathbb{R}^n$. Для линейных операторов задается отношение частичного порядка на конусе неотрицательно определенных матриц следующим стандартным образом:

$$A \preceq A_1, \quad A_1 \succeq A, \quad A\colon \mathbb{R}^n \to \mathbb{R}^n, \quad A_1\colon \mathbb{R}^n \to \mathbb{R}^n \quad \Leftrightarrow \quad \langle (A_1 - A)x, x\rangle \geqslant 0 \quad \forall x \in \mathbb{R}^n.$$

Аналогичное отношение верно и относительно сопряженного пространства:

$$B \preceq B_1, \quad B_1 \succeq B, \quad B\colon \mathbb{R}^m \to \mathbb{R}^m, \quad B_1\colon \mathbb{R}^m \to \mathbb{R}^m \quad \Leftrightarrow \quad \langle u, (B_1 - B)u\rangle \geqslant 0 \quad \forall u \in \mathbb{R}^m.$$

Заметим, что для линейного оператора $A\colon \mathbb{R}^n \to \mathbb{R}^m$ верны отношения

$$\begin{cases} AA^\top \succeq \sigma_{\min}\left(A^\top\right)^2 I_m, \\ A^\top A \succeq \sigma_{\min}(A)^2 I_n. \end{cases}$$

Обозначим через $\overline{1, m}$ множество целых чисел от 1 до m включительно. Обозначим через $f(x) = \mathrm{O}(h(x))$ оценку сверху функции f функцией h с точностью до константы. Положим также

$$f^* \overset{\text{def}}{=} \min_{x\in\mathbb{R}^n} f(x),$$

определив минимальное возможное значение по аргументу x для функции f.

Регуляризация метода Гаусса – Ньютона

Линеаризация и мажорирование оптимизируемой функции

Рассмотрим итеративную процедуру решения задачи (2), основанную на минимизации *линеаризованной модели* оптимизируемой функции:

$$\phi(x, y) \stackrel{\text{def}}{=} \|F(x) + F'(x)(y - x)\|, \quad (x, y) \in \mathbb{R}^n \times \mathbb{R}^n.$$

Дополнительно определим условия решаемой задачи. Рассмотрим замкнутое выпуклое множество $\mathcal{F} \subseteq \mathbb{R}^n$ с непустой внутренностью.

Предположение 1. *Пусть $F(x)$ — векторнозначное отображение, удовлетворяющее условию Липшица на $\mathcal{F}$:*

$$\exists L_F > 0\colon \|F'(y) - F'(x)\| \leqslant L_F\|y - x\| \quad \forall (x, y) \in \mathcal{F}^2.$$

Введем понятие множества Лебега уровня $f_1(x_k)$ функции f_1 относительно приближенного решения x_k:

$$\mathcal{L}(f_1(x_k)) \stackrel{\text{def}}{=} \{x\colon f_1(x) \leqslant f_1(x_k)\}.$$

Предположим $\mathcal{L}(f_1(x_0)) \subseteq \mathcal{F}$, где $x_0 \in \mathcal{F}$ — начальное приближение.

Предположение 2. *Пусть для многозначного отображения выполнено неравенство Поляка – Лоясиевича* [Поляк, 1963]*, то есть существует такой $\mu > 0$, для которого при любых $x \in \mathcal{F}$ выполнено неравенство*

$$\sigma_{\min}\left(F'(x)^\top\right) \geqslant \sqrt{\mu}. \tag{3}$$

В [Nesterov, 2020] предложена следующая локальная мажоранта (локальная модель) оптимизируемой функции $f_1(y)$:

$$f_1(y) \leqslant \frac{f_1(x)}{2} + \frac{(\phi(x, y))^2}{2f_1(x)} + \frac{L}{2}\|y - x\|^2, \quad L \geqslant L_F, \quad (x, y) \in \mathcal{F}^2. \tag{4}$$

В текущей работе рассматривается общая форма мажоранты, ее вывод изложен в лемме 1 (формулировка и доказательство представлены в приложении):

$$f_1(y) \leqslant \psi_{x,L,\tau}(y) \stackrel{\text{def}}{=} \frac{\tau}{2} + \frac{(\phi(x, y))^2}{2\tau} + \frac{L}{2}\|y - x\|^2, \quad L \geqslant L_F, \quad \tau > 0, \quad (x, y) \in \mathcal{F}^2.$$

Приведенная мажоранта позволяет определить формулу обновления приближенного решения в итерационной схеме:

$$T_{L,\tau}(x) \stackrel{\text{def}}{=} \arg\min_{y \in \mathbb{R}^n}\{\psi_{x,L,\tau}(y)\} = x - \left(F'(x)^\top F'(x) + \tau L I_n\right)^{-1} F'(x)^\top F(x).$$

Схема метода

Предлагаемая схема метода представлена в виде алгоритма 1. Эта схема является ускоренной с помощью моментного члена в пункте 7, при этом описание алгоритма позволяет на практике решать задачи без моментного члена, положив $t_k = 0$. Сама схема может осуществлять неточное вычисление промежуточной точки y_{k+1}, ограничивая неточность вычисления с помощью значения ε_k. В схеме также возможно варьирование значения τ_k, что позволяет на практике получать различные версии метода, например, $\tau_k = f_1(x_k)$ соответствует ускоренной версии метода трех квадратов.

Алгоритм 1. Ускоренный метод трех квадратов с неточным проксимальным отображением

Input:

$$\begin{cases} y_0 = x_0 \in \mathbb{R}^n,\ \mathcal{L}(f_1(x_0)) \subseteq \mathcal{F} \text{ — начальное приближение, } x_{-1} = x_0, \\ \mathcal{E}(\cdot) \text{ — функция погрешности вычисления } y_{k+1}, \\ N \in \mathbb{N} \text{ — количество итераций метода,} \\ L \text{ — оценка локальной постоянной Липшица, } L > 0,\ L_0 = L, \\ \mathcal{T}(\cdot) \text{ — функция, определяющая значение } \tau. \end{cases}$$

for $k = 0, 1, \ldots, N-1$ **do**

1) определить $\tau_k = \mathcal{T}(x_k, L_k, \varepsilon_k)$, $\varepsilon_k = \mathcal{E}(k, x_k, x_{k-1})$;
2) вычислить такой $y_{k+1} \in \mathbb{R}^n$, что $\psi_{x_k,L_k,\tau_k}(y_{k+1}) - \psi_{x_k,L_k,\tau_k}\left(T_{L_k,\tau_k}(x_k)\right) \leqslant \varepsilon_k$ и $f_1(x_k) - \psi_{x_k,L_k,\tau_k}(y_{k+1}) \geqslant 0$;
3) если $f_1(y_{k+1}) > \psi_{x_k,L_k,\tau_k}(y_{k+1})$, то положить $L_k := 2L_k$ и вернуться к пункту 1;
4) $L_{k+1} = \max\left\{\frac{L_k}{2}, L\right\}$;
5) определить $\varphi_k(t) = f_1(y_{k+1} + t(y_{k+1} - y_k))$;
6) найти такой $t_k \geqslant 0$, что $\varphi_k(t_k) \leqslant \varphi_k(0)$;
7) положить $x_{k+1} = y_{k+1} + t_k(y_{k+1} - y_k)$.

end for

Output: x_N.

На каждой итерации метода происходит поиск промежуточной точки y_{k+1}, при этом требуется выполнение условия гарантированной монотонности:

$$f_1(x_k) - \psi_{x_k,L_k,\tau_k}(y_{k+1}) \geqslant 0,$$

которое, например, автоматически выполнено в случае $\tau_k = f_1(x_k)$ и $\varepsilon_k = 0$ [Nesterov, 2020]. Схема не требует знания постоянной Липшица, в ней реализована процедура поиска локальной постоянной L_k. Также в схему включен поиск фактора моментного ускорения t_k с условием $\varphi_k(t_k) \leqslant \varphi_k(0)$, позволяющим анализировать сходимость схемы, используя только последовательность $\{x_k\}_{k\in\mathbb{Z}_+}$ в силу монотонности по значению функции:

$$f_1(x_k) \geqslant f_1(y_{k+1}) \geqslant f_1(x_{k+1}).$$

Алгоритм 1 содержит в себе моментное ускорение общего вида, в формате «черного ящика», с требованием монотонности последовательности $\{f_1(x_k)\}_{k\in\mathbb{Z}_+}$ при вычислении $t_k \geqslant 0$.

Анализ сходимости

В итерационном процессе правило вида $x_{k+1} = T_{L_k,\tau_k}(x_k)$ при выполнении дополнительных условий невырожденности позволяет получить глобальную линейную сходимость для произвольного начального приближения $x_0 \in \mathcal{F}$, $\mathcal{L}(f_1(x_0)) \subseteq \mathcal{F}$ в случае $\tau_k = f_1(x_k)$ и $\varepsilon_k = 0$ [Nesterov, 2020]. В данной работе рассматривается более общая форма правила обновления x_{k+1}, использующая явное выражение $T_{L_k,\tau_k}(x_k)$:

$$y_{k+1} = y(\eta_k) \overset{\text{def}}{=} x_k - \eta_k \left(F'(x_k)^\top F'(x_k) + \tau_k L_k I_n\right)^{-1} F'(x_k)^\top F(x_k), \quad \eta_k \geqslant 0. \tag{5}$$

Полученное таким образом значение y_{k+1} является субоптимальным относительно $T_{L_k,\tau_k}(x_k)$, однако позволяет лучше рассмотреть метод Гаусса – Ньютона в классе квазиньютоновских методов. На практике такие субоптимальные шаги могут привести к более лучшему убыванию значения минимизируемой функции f_1, например, с помощью произвольной процедуры линейного поиска по η:

$$\eta_k \in \operatorname*{Arg\,min}_{\eta \geqslant 0} \left\{ f_1 \left(x_k - \eta \left(F'(x_k)^\top F'(x_k) + \tau_k L_k I_n\right)^{-1} F'(x_k)^\top F(x_k) \right) \right\} \quad \text{или}$$

$$\eta_k \in \operatorname*{Arg\,min}_{\eta \in [0,\, 1]} \left\{ f_1 \left(x_k - \eta \left(F'(x_k)^\top F'(x_k) + \tau_k L_k I_n\right)^{-1} F'(x_k)^\top F(x_k) \right) \right\}.$$

То есть процедура линейного поиска вносит дополнительную свободу для уменьшения $f_1(y_{k+1})$ по сравнению с оригинальным методом трех квадратов, в котором полагают $\eta_k = 1$. В результате использования линейного поиска по η выполняется цепочка нестрогих неравенств ниже:

$$f_1(y(\eta_k)) \leqslant f_1(y(1)) \leqslant f_1(x_k)$$

при условии выбора достаточно большого значения L_k. Следующее утверждение определяет множество оценок на значение $f_1(x_{k+1})$.

Теорема 1. *Пусть выполнены предположения* 1 *и* 2*. Рассмотрим последовательность* $\{x_k\}_{k \in \mathbb{Z}_+}$*, вычисляемую по схеме* 1 *с правилом* (5)*, в котором* $\tau_k > 0$*,* $\eta_k \in (0, 2)$*,* $k \in \mathbb{Z}_+$*. Тогда для* $k \in \mathbb{Z}_+$*:*

$$f_1(x_{k+1}) \leqslant \frac{\tau_k}{2} + \left[\begin{array}{l} \dfrac{f_2(x_k)}{2\tau_k} \left(1 - \dfrac{\eta_k(2-\eta_k)\mu}{L_k\tau_k + \mu}\right) \quad \textit{при } \tau_k \geqslant \dfrac{\mu}{L_k}, \\ \dfrac{f_2(x_k)(1-\eta_k)^2}{2\tau_k} + \dfrac{\eta_k(2-\eta_k)L_k f_2(x_k)}{2\mu} - \dfrac{\eta_k(2-\eta_k)L_k^2 f_2(x_k)\tau_k}{2\mu^2(1+\xi)^3} \quad \textit{для} \\ \textit{некоторого } \xi \in (-1,\, 1] \textit{ при } \tau_k < \dfrac{\mu}{L_k}. \end{array} \right.$$

При этом для $\eta_k = 1$*,* $k \in \mathbb{Z}_+$*, верно*

$$f_1(x_{k+1}) \leqslant \frac{\tau_k}{2} + \frac{L_F}{\mu} f_2(x_k).$$

Использование явной формы $T_{L_k,\tau_k}(x_k)$ в анализе позволяет устанавливать сходимость относительно градиента функции f_2, как указано в утверждении ниже.

Теорема 2. *Пусть выполнено предположение* 1. *Рассмотрим последовательность* $\{x_k\}_{k\in\mathbb{Z}_+}$, *вычисляемую по схеме* 1 *с правилом* (5), *в котором* $\tau_k > 0$ *и* $\eta_k > 0$, $\eta_k(2 - \eta_k) \geqslant c > 0$, $k \in \mathbb{Z}_+$. *Дополнительно предположим ограниченность нормы матрицы Якоби: существует* $M_F > 0$, *для которого выполнено* $\|F'(x)\| \leqslant M_F$ *при всех* $x \in \mathcal{F}$. *Тогда при* $k \in \mathbb{N}$

$$\min_{i\in\overline{0,\,k-1}}\left\{\|\nabla f_2(x_i)\|^2\right\} \leqslant \frac{8\left(2L_F\max\limits_{i\in\overline{0,\,k-1}}\{\tau_i\} + M_F^2\right)}{\eta(2-\eta)k}\sum_{i=0}^{k-1}\left(\frac{1}{2}\left(\tau_i - f_1(x_i)\right)^2 + \tau_i(f_1(x_i) - f_1(x_{i+1}))\right),$$

$$\eta \in \operatorname*{Arg\,min}_{k\in\mathbb{Z}_+}\{\eta_k(2-\eta_k)\}.$$

Выбор $\tau_k = f_1(x_k)$ в теореме 2 приводит к упрощению оценки на квадрат нормы градиента в силу наличия сходимости ($f_1(x_k)f_1(x_{k+1}) \geqslant f_2(x_{k+1})$, $k \in \mathbb{Z}_+$):

$$\min_{i\in\overline{0,\,k-1}}\left\{\|\nabla f_2(x_i)\|^2\right\} \leqslant \frac{8f_2(x_0)\left(2L_Ff_1(x_0) + M_F^2\right)}{\eta(2-\eta)k}.$$

Для достижения минимального значения нормы градиента f_2 на уровне $\epsilon > 0$ необходимо затратить $k = \mathrm{O}\left(\frac{1}{\epsilon^2}\right)$ итераций метода:

$$k = \left\lceil \frac{8f_2(x_0)\left(2L_Ff_1(x_0) + M_F^2\right)}{\eta(2-\eta)\epsilon^2}\right\rceil,$$

формально для такого количества итераций выполнено неравенство

$$\min_{i\in\overline{0,\,k-1}}\left\{\|\nabla f_2(x_i)\|\right\} \leqslant \epsilon. \tag{6}$$

Представленные теоремы 1 и 2 полезны для общего понимания того, как устроена процедура оптимизации в методе трех квадратов, однако далеко не любая последовательность $\{\tau_k\}_{k\in\mathbb{Z}_+}$ обеспечивает гарантированную сходимость метода даже к стационарной точке. С целью внести дополнительную ясность в определение τ_k рассмотрим следующее утверждение.

Теорема 3. *Пусть условия теоремы* 1 *верны, определим на каждой итерации метода трех квадратов максимальный коэффициент линейной сходимости* α_k:

$$\alpha_k \in \left[\frac{1}{2} + \frac{L_kf_1(x_k) + (1-\eta_k)^2\mu}{2(L_kf_1(x_k) + \mu)},\ 1\right), \quad k \in \mathbb{Z}_+.$$

Дополнительно предположим, что на каждой итерации $\tau_k = c_kf_1(x_k)$, $k \in \mathbb{Z}_+$:

$$c_k \in \left[\frac{L_kf_1(x_k) + (1-\eta_k)^2\mu}{(2\alpha_k - 1)(L_kf_1(x_k) + \mu)},\ \alpha_k - \frac{1}{2} - \frac{\mu}{2L_kf_1(x_k)} + \right.$$
$$\left. + \sqrt{\alpha_k^2 + \alpha_k\left(\frac{\mu}{L_kf_1(x_k)} - 1\right) + \frac{1}{4}\left(\frac{\mu}{L_kf_1(x_k)} + 1\right)^2 - \frac{(1-\eta_k)^2\mu}{L_kf_1(x_k)}}\ \right].$$

Тогда метод трех квадратов с вычислением x_{k+1} *по правилу* (5) *глобально сходится не хуже, чем линейно, к решению задачи* (2) $\lim\limits_{k\to+\infty} x_k = x^*: F(x^*) = \mathbf{0}_m$ *со следующей оценкой*:

$$f_1(x_k) \leqslant f_1(x_0)\prod_{i=0}^{k-1}\alpha_i, \quad k \in \mathbb{Z}_+, \quad \prod_{i=0}^{-1}\alpha_i \overset{\mathrm{def}}{=} 1.$$

Утверждение выше определяет возможные значения τ_k, пропорциональные $f_1(x_k)$, при которых метод трех квадратов имеет глобально линейную сходимость. Стоит заметить, что наименьшее допустимое значение α_k соответствует $c_k = 1$. Это представляет метод трех квадратов с $\tau_k = f_1(x_k)$ как наиболее быстрый в классе методов трех квадратов с гарантией линейной сходимости, отличающихся выбором c_k. Данный результат является скорее теоретико-иллюстративным, обосновывающим важность и общее удобство случая $\tau_k = f_1(x_k)$. Также в условии теоремы 3 роль η_k характерно показана: близость значений η_k к 0 и к 2 относительно $\eta_k = 1$ симметрично приводит к замедлению сходимости по значению α_k. В теореме 3 выбор $\tau_k = f_1(x_k)$ соответствует следующей оценке:

$$\begin{aligned} f_1(x_{k+1}) &\leqslant f_1(x_k)\left(1 - \frac{\eta_k(2-\eta_k)\mu}{2(L_k f_1(x_k) + \mu)}\right) \leqslant f_1(x_k)\left(1 - \frac{\eta(2-\eta)\mu}{2(2L_F f_1(x_0) + \mu)}\right) \leqslant \\ &\leqslant f_1(x_k)\exp\left(-\frac{\eta(2-\eta)\mu}{2(2L_F f_1(x_0) + \mu)}\right) \leqslant \\ &\leqslant f_1(x_0)\exp\left(-\frac{(k+1)\eta(2-\eta)\mu}{2(2L_F f_1(x_0) + \mu)}\right), \quad \eta \in \operatorname*{Arg\,min}_{k \in \mathbb{Z}_+}\{\eta_k(2-\eta_k)\}. \end{aligned}$$

Благодаря строению функции f_2 возможно установить аналогичную оценку на норму градиента:

$$\|\nabla f_2(x_k)\| = \left\|2F'(x_k)^\top F(x_k)\right\| \leqslant 2\left\|F'(x_k)^\top\right\| \cdot \|F(x_k)\| = 2\|F'(x_k)\| f_1(x_k).$$

Оценки в теоремах 1 и 3 при $\tau_k = f_1(x_k)$ означают для достижения уровня

$$f_1(x_k) \leqslant \epsilon \tag{7}$$

необходимость затратить $k = \mathrm{O}\left(\ln\left(\frac{1}{\epsilon}\right)\right)$ итераций:

$$k = \left\lceil \frac{2(2L_F f_1(x_0) + \mu)}{\eta(2-\eta)\mu} \ln\left(\frac{f_1(x_0)}{\epsilon}\right)\right\rceil.$$

Таким образом, теорема 3 устанавливает глобальную линейную сходимость при выполнении условия Поляка – Лоясиевича с гарантией отсутствия участков сублинейной сходимости.

Эксперименты

На основе разработанного алгоритма 1 проведена серия экспериментов для практической верификации производительности в случае использования точного оракула (5) с $\tau_k = f_1(x_k)$, $k \in \mathbb{Z}_+$, в отсутствие и при наличии линейного поиска по η. Программный код для выполнения экспериментов на языке программирования Python 3.8 выложен в открытый репозиторий[1].

Метод протестирован на двух задачах безусловной минимизации функции f посредством поиска стационарной точки:

$$\nabla f(x) = \mathbf{0}_n.$$

В рассмотренных задачах функции невыпуклы и как минимум дважды дифференцируемы. Использованные функции $f(x)$, $x^\top \overset{\text{def}}{=} \left(x^1, \ldots, x^n\right)$, $x \in \mathbb{R}^n$, перечислены ниже.

[1] https://github.com/neyudin/AcceleratedGNMethodEquations

- Функция Розенброка – Скокова [Гасников, 2020]:

$$f_{\mathrm{RS}}(x) = \sum_{i=1}^{n-1} \left(i^2 \left(x^i - \left(x^{i+1} \right)^2 \right)^2 + \left(1 - x^{i+1} \right)^2 \right).$$

При этом для f_{RS} $m = 2n - 2$ и F задается набором функций, $j \in \overline{1, m}$:

$$F_j(x) = \begin{cases} 1 - x^{(j+2)/2}, & \text{если } j \vdots 2; \\ \left(\dfrac{j+1}{2} \right) \left(x^{(j+1)/2} - \left(x^{(j+3)/2} \right)^2 \right), & \text{если иначе.} \end{cases}$$

- Hat-функция: $f_{\mathrm{H}}(x) = \left(\|x\|^2 - 1 \right)^2$ с $F(x) = \nabla f_{\mathrm{H}}(x)$.

Значение размерности задачи n равно 100. Сходимость алгоритма 1 оценивается по величине функции f_1 и относительно нормы градиента функции f_2.

В реализацию алгоритма 1 добавлен критерий предварительного останова на каждой итерации k по норме градиента и по значению τ, выражающийся в форме предиката

$$(\tau_k < \varepsilon) \quad \text{или} \quad (\|\nabla f_2(x_k)\| < \varepsilon),$$

где $\varepsilon = 10^{-6}$. На каждом наборе данных работа алгоритма 1 длится не дольше $N = 1000$ внешних итераций. В алгоритме 1 процедура поиска подходящего t_k реализована следующими способами.

- *Константная стратегия* (*неускоренный метод*), соответствует значению $t_k = 0$, то есть отсутствию моментного ускорения.

- *Стратегия экстраполяции* [Nesterov, 2020], состоит в циклическом переборе значений $t_k \geqslant 0$, начиная с $t_k = 1$, с умножением текущего t_k на 2 для получения нового значения t_k, пока субградиент $\varphi_k'(t_k)$ не станет неотрицательным или пока не произойдет нарушение монотонного невозрастания $\varphi_k(t_k)$ между соседними итерациями цикла. В самом начале монотонность проверяется по условию $\varphi_k(0) \geqslant \varphi_k(1)$. Если останов цикла произошел в самом начале, то присваивается $t_k = 0$.

- *Правило Армихо* [Nesterov, 2020; Armijo, 1966; Wolfe, 1969; Wolfe, 1971], заключается в поиске на луче $t_k \geqslant 0$ такого значения t_k, при котором выполнены следующие неравенства: $\varphi_k(0) + c_2 \varphi_k'(0) t_k \leqslant \varphi_k(t_k) \leqslant \varphi_k(0) + c_1 \varphi_k'(0) t_k$ для фиксированных $0 < c_1 < c_2 < 1$. Если при этом субградиент $\varphi_k'(0) \geqslant 0$, то выбирается $t_k = 0$.

Линейный поиск η осуществлялся по правилу Армихо для функции $f_1(y(\eta))$. Результаты экспериментов отображены в графиках зависимости $f_1(x_k)$ от итерации k и зависимости $\|\nabla f_2(x_k)\|$ от итерации k. Графики представлены в виде усреднений по пяти запускам алгоритма 1, каждый запуск отличается начальным приближением x_0, полученным сэмплированием из стандартного многомерного нормального распределения. Дополнительно на каждом рисунке проведена горизонтальная линия $\varepsilon = 10^{-6}$, обозначающая уровень предварительного останова.

При сравнении результатов неускоренного метода на рис. 1 (представлен в приложении) с ускоренными аналогами можно утверждать, что все предложенные стратегии позволяют в разной степени уменьшить количество необходимых итераций до сходимости. Точно можно сказать, что наименьший прирост в скорости сходимости при использовании моментного члена демонстрирует алгоритм 1 на функции Розенброка – Скокова f_{RS}, причем сходимость всегда сублинейная. Заметнее эффект от ускорения наблюдается на функции f_{H}, сходимость при непостоянном t_k, в свою очередь, типично линейная и даже суперлинейная, судя по представленным графикам на каждом рисунке. Эксперименты показали, что на функции f_{RS} важно само наличие линейного поиска η, результат работы алгоритма 1 слабо зависит от гиперпараметров линейного поиска по Армихо. В случае $t_k \equiv 0$ такое утверждение верно и для f_H, причем при использовании стратегии экстраполяции линейный поиск η для f_H только замедлял сходимость. Если же сравнить между собой правило Армихо и стратегию экстраполяции для определения t_k, то в случае функции f_{RS} эффективнее относительно затраченных итераций показало себя правило Армихо, однако в случае функции f_H такое можно утверждать только в адрес нескольких значений (c_1, c_2), задающих гиперпараметры правила Армихо.

Заключение

В работе предложена схема метода Гаусса – Ньютона с доказанной верхней границей скорости сходимости в худшем случае, в схеме 1 присутствует предложенное в этой работе моментное ускорение произвольной формы, не замедляющее работу алгоритма 1 в худшем случае относительно количества затраченных внешних итераций, причем оценка носит глобальный характер. При этом введенный моментный член позволяет ускорить работу метода на практике, что экспериментально продемонстрировано с помощью разработанных в рамках данной работы стратегий вычисления подходящего моментного члена. Также заметное улучшение сходимости на практике добавляет использование линейного поиска шага метода η. По своей форме предложенный алгоритм 1 напоминает алгоритм Левенберга – Марквардта [Levenberg, 1944; Marquardt, 1963; Moré, 1978], при этом алгоритм 1 имеет такую же сложность в худшем случае, что и алгоритм Левенберга – Марквардта при аналогичных условиях [Zhao, Fan, 2016; Marumo et al., 2020]. Но, в отличие от представленных в [Zhao, Fan, 2016; Marumo et al., 2020] версий алгоритма Левенберга – Марквардта, предложенный алгоритм 1 легче реализуем на практике, не требует многовариантных ветвящихся схем, характерных для методов доверительной области. Более того, изложенный метод Гаусса – Ньютона позволяет изменить мажоранту в алгоритме 1 и обобщить ускоренную схему на случай неевклидовых норм интерпретируемым образом.

Список литературы (References)

Гасников А. В. Современные численные методы оптимизации. Метод универсального градиентного спуска // e-print, 2020. — https://arxiv.org/pdf/1711.00394.pdf

Gasnikov A. V. Sovremennye chislennye metody optimizatsii. Metod universal'nogo gradientnogo spuska [Universal gradient descent] // e-print, 2020. — https://arxiv.org/pdf/1711.00394.pdf (in Russian).

Голиков А. И., Евтушенко Ю. Г., Капорин И. Е. Метод ньютоновского типа для решения систем линейных уравнений и неравенств // Журнал вычислительной математики и математической физики. — 2019. — Т. 59, № 12. — С. 2086–2101.

Golikov A. I., Evtushenko Yu. G., Kaporin I. E. Metod n'yutonovskogo tipa dlya resheniya sistem lineinykh uravnenii i neravenstv [Newton-type method for solving systems of linear equations and inequalities] // Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki. — 2019. — Vol. 59, No. 12. — P. 2086–2101 (in Russian).

Поляк Б. Т. Градиентные методы минимизации функционалов // Журнал вычислительной математики и математической физики. — 1963. — Т. 3, № 4. — С. 643–653.

Polyak B. T. Gradientnye metody minimizacii funkcionalov [Gradient methods for minimizing functionals] // Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki. — 1963. — Vol. 3, No. 4. — P. 643–653 (in Russian).

Armijo L. Minimization of functions having Lipschitz continuous first partial derivatives // Pacific Journal of mathematics. — 1966. — Vol. 16, No. 1. — P. 1–3.

Avriel M., Wilde D. J. Optimally proof for the symmetric Fibonacci search technique // Fibonacci Quarterly Journal. — 1966.

Botev A., Ritter H., Barber D. Practical Gauss – Newton optimisation for deep learning // International Conference on Machine Learning. — PMLR, 2017. — P. 557–565.

Cai T., Gao R., Hou J., Chen S., Wang D., He D., Zhang Z., Wang L. Gram – Gauss – Newton method: learning overparameterized neural networks for regression problems // arXiv preprint. — 2019. — arXiv:1905.11675

Kiefer J. Sequential minimax search for a maximum // Proceedings of the American mathematical society. — 1953. — Vol. 4, No. 3. — P. 502–506.

Le H., Zach C., Rosten E., Woodford O. J. Progressive batching for efficient non-linear least squares // Proceedings of the Asian Conference on Computer Vision. — 2020.

Levenberg K. A method for the solution of certain non-linear problems in least squares // Quarterly of applied mathematics. — 1944. — Vol. 2, No. 2. — P. 164–168.

Marquardt D. W. An algorithm for least-squares estimation of nonlinear parameters // Journal of the society for Industrial and Applied Mathematics. — 1963. — Vol. 11, No. 2. — P. 431–441.

Marumo N., Okuno T., Takeda A. Constrained Levenberg – Marquardt method with global complexity bound // arXiv preprint. — 2020. — arXiv:2004.08259

Moré J. J. The Levenberg – Marquardt algorithm: implementation and theory // Numerical analysis. — Berlin, Heidelberg: Springer, 1978. — P. 105–116.

Nesterov Yu. Flexible modification of Gauss – Newton method // CORE Discussion Papers. — 2021.

Nesterov Yu. Lectures on convex optimization. — Vol. 137. — Berlin, Germany: Springer, 2018.

Nesterov Yu. Modified Gauss – Newton scheme with worst case guarantees for global performance // Optimisation methods and software. — 2007. — Vol. 22, No. 3. — P. 469–483.

Ren Y., Goldfarb D. Efficient subsampled Gauss – Newton and natural gradient methods for training neural networks // arXiv preprint. — 2019. — arXiv:1906.02353

Thiele C., Araya-Polo M., Hohl D. Deep neural network learning with second-order optimizers — a practical study with a stochastic quasi-Gauss – Newton method // arXiv preprint. — 2020. — arXiv:2004.03040

Wolfe P. Convergence conditions for ascent methods // SIAM review. — 1969. — Vol. 11, No. 2. — P. 226–235.

Wolfe P. Convergence conditions for ascent methods. II: Some corrections // SIAM review. — 1971. — Vol. 13, No. 2. — P. 185–188.

Zhao R., Fan J. Global complexity bound of the Levenberg – Marquardt method // Optimization Methods and Software. — 2016. — Vol. 31, No. 4. — P. 805–814.