ТРУДЫ ИНСТИТУТА МАТЕМАТИКИ И МЕХАНИКИ УрО РАН

Том 25 № 4 2019

УДК 519.85

АДАПТАЦИЯ К ВЕЛИЧИНАМ ПОГРЕШНОСТЕЙ ДЛЯ НЕКОТОРЫХ МЕТОДОВ ОПТИМИЗАЦИИ ГРАДИЕНТНОГО ТИПА[1]

Ф. С. Стонякин

Введено новое понятие неточной модели выпуклой целевой функции, учитывающее возможность погрешностей при как задании функции, так и ее градиента. Для этой концепции предложен градиентный метод с адаптивной настройкой некоторых параметров модели, и получена оценка скорости сходимости. Эта оценка оптимальна на классе достаточно гладких задач при наличии погрешностей. Рассмотрен специальный класс задач выпуклой негладкой оптимизации, к которым применима предложенная методика за счет искусственного введения неточности. Показано, что для таких задач возможно модифицировать метод так, чтобы гарантированно имела место сходимость по функции со скоростью, близкой к оптимальной на классе задач выпуклой негладкой оптимизации. Предложен адаптивный градиентный метод для целевых функций с некоторой релаксацией условия липшицевости градиента, удовлетворяющих условию градиентного доминирования Поляка — Лоясиевича. При этом учитывается возможность неточного задания целевой функции и градиента. Адаптивный выбор параметров при работе метода выполняется как по величине константы Липшица градиента, так и по величине, соответствующей погрешности задания градиента и целевой функции. Обоснована линейная сходимость метода с точностью до величины, связанной с погрешностями.

Ключевые слова: градиентный метод, адаптивный метод, липшицев градиент, негладкая оптимизация, условие градиентного доминирования.

F. S. Stonyakin. Adaptation to inexactness for some gradient-type optimization methods.

We introduce a notion of inexact model of a convex objective function, which allows for errors both in the function and in its gradient. For this situation, a gradient method with an adaptive adjustment of some parameters of the model is proposed and an estimate for the convergence rate is found. This estimate is optimal on a class of sufficiently smooth problems in the presence of errors. We consider a special class of convex nonsmooth optimization problems. In order to apply the proposed technique to this class, an artificial error should be introduced. We show that the method can be modified for such problems to guarantee a convergence in the function with a nearly optimal rate on the class of convex nonsmooth optimization problems. An adaptive gradient method is proposed for objective functions with some relaxation of the Lipschitz condition for the gradient that satisfy the Polyak–Lojasievicz gradient dominance condition. Here, the objective function and its gradient can be given inexactly. The adaptive choice of the parameters is performed during the operation of the method with respect to both the Lipschitz constant of the gradient and a value corresponding to the error of the gradient and the objective function. The linear convergence of the method is justified up to a value associated with the errors.

Keywords: gradient method, adaptive method, Lipschitz gradient, nonsmooth optimization, gradient dominance condition.

MSC: 90C25, 90C06, 65K10

DOI: 10.21538/0134-4889-2019-25-4-210-225

1. Введение

Хорошо известно, что методы градиентного типа отличаются относительной простотой и малыми затратами памяти, что объясняет их популярность в работах по многомерной оптимизации (см., например, [1–9]). Напомним, что для вывода оценок скорости сходимости градиентного метода можно использовать идею аппроксимации функции в исходной точке (текущем положении метода) мажорирующим ее параболоидом вращения. Так, для задачи минимизации выпуклого функционала $f : Q \to \mathbb{R}$, заданного на выпуклом замкнутом множестве $Q \subset \mathbb{R}^n$ с

[1]Работа выполнена при поддержке Российского научного фонда, проект 18-71-00048.

липшицевым градиентом (для некоторой константы $L > 0$ при произвольных $x, y \in Q$ верно $\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|$), выполняются неравенства

$$f(x) + \langle \nabla f(x), y - x\rangle \leqslant f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2. \tag{1.1}$$

Неравенства (1.1) позволяют получить для обычного градиентного спуска с постоянным шагом оценку скорости сходимости

$$f(\hat{x}) - f^* \leqslant \frac{C_1}{N}, \tag{1.2}$$

где $\hat{x}$ — выход работы метода после N итераций, f^* — точное значение искомого минимума функции f, C_1 — некоторая положительная константа.

В новых работах, посвященных методам градиентного типа, например в [4], введено условие относительной гладкости оптимизируемого функционала, предполагающее замену правого неравенства в (1.1) на ослабленный вариант

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + LV(y, x),$$

где $V(y, x)$ — широко используемый в оптимизации аналог расстояния между точками x и y, который называют *расхождением Брэгмана*. Обычно *расхождение Брэгмана* вводится с использованием вспомогательной 1-сильно выпуклой функции d (порождает расстояния), которая дифференцируема во всех точках $x \in Q$,

$$V(y, x) = d(x) - d(y) - \langle \nabla d(x), y - x\rangle \quad \forall x, y \in Q;$$

здесь $\langle \cdot, \cdot \rangle$ — скалярное произведение в $\mathbb{R}^n$, причем ввиду 1-сильной выпуклости функции d для произвольных $x, y \in Q$ верно неравенство $V(y, x) \geqslant 1/2\,\|y - x\|^2$. В частности, для стандартной евклидовой нормы $\|\cdot\|_2$ и соответствующего расстояния в $\mathbb{R}^n$ можно считать, что $V(y, x) = d(y, x) = 1/2\,\|y - x\|_2^2$ для произвольных $x, y \in Q$. Однако рассмотренное в [4] условие относительной гладкости предполагает лишь выпуклость (но не сильную выпуклость) порождающей функции d. Как показано в [4], концепция относительной гладкости позволяет применить вариант градиентного метода для некоторых задач, к которым ранее применялись лишь методы внутренней точки.

Весьма естественно возникает вопрос влияния на скорость сходимости метода погрешностей задания целевой функции и/или градиента. В этом плане можно отметить хорошо известную концепцию неточного оракула О. Деволдера — Ф. Глинера — Ю. Е. Нестерова [5; 6]. Говорят, что функция f допускает неточный оракул $(f_\delta(x), g_\delta(x)) \in \mathbb{R} \times \mathbb{R}^n$ в произвольной запрошенной точке $x \in Q$, если для некоторых $\delta > 0$ и $L > 0$ выполняется аналог неравенства (1.1):

$$f_\delta(x) + \langle g_\delta(x), y - x\rangle \leqslant f(y) \leqslant f_\delta(x) + \langle g_\delta(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 + \delta \ \ \forall x, y \in Q. \tag{1.3}$$

По сути, (1.3) означает, что $f_\delta(x)$ — приближенное значение $f(x)$ ($f_\delta(x) \leqslant f(x) \leqslant f_\delta(x) + \delta$), а $g_\delta(x)$ — некоторый δ-субградиент f в произвольной точке x. Оказывается [5], что при выполнении условия (1.3) для градиентного метода (с заменой пары $(f, \nabla f)$ на (f_δ, g_δ)) верна оценка скорости сходимости

$$f(\hat{x}) - f^* \leqslant \frac{C_1}{N} + \delta,$$

т. е. в оценке не происходит накопления величин, соответствующих погрешностям.

Идеология Деволдера — Глинера — Нестерова была развита в работе [2], где обобщена концепция (δ, L)-оракула и введено понятие (δ, L)-модели целевой функции. Идея этого обобщения заключается в том, что линейная функция $\langle \nabla f(y), x - y\rangle$ в неравенстве (1.3) заменяется на некоторую абстрактную выпуклую функцию $\psi(x, y)$ [2].

О п р е д е л е н и е 1. Говорят, что функция f допускает (δ, L)-модель $(f_\delta(x), \psi(y,x))$ в точке $x \in Q$, если для любого $y \in Q$ справедливо неравенство

$$0 \leq f(y) - f_\delta(x) - \psi(y,x) \leq \frac{L}{2}\|y-x\|^2 + \delta, \tag{1.4}$$

где $\psi(x,x) = 0 \ \forall x \in Q$ и $\psi(x,y)$ — выпуклая функция по x для всякого $y \in Q$.

Концепция из определения 1 позволяет обосновать сходимость градиентного метода для достаточно широкого класса задач оптимизации [2;3]. По сути, она дает возможность унифицировать подходы к различным на первый взгляд классам задач оптимизации с описанием степени влияния погрешностей данных на гарантированное качество решения, достижимое в ходе работы метода.

В настоящей работе предлагается модификация концепции (δ, L)-модели целевой функции, которая учитывает возможность неточного задания не только значения целевой функции, но и самой функции-модели. В частности, для линейной модели $\psi(x,y) = \langle \nabla f(x), y-x\rangle$ описывается ситуация некоторой модификации условий (1.4) с учетом отдельно погрешности задания f и ∇f. Если положить, что $\forall x \in Q$ справедливо

$$\left\|\nabla f(x) - \widetilde{\nabla} f(x)\right\| \leqslant \Delta, \quad \Delta > 0 \tag{1.5}$$

для некоторого доступного приближенного значения $\widetilde{\nabla} f(x)$ градиента ∇f, то будет верно неравенство $\left|\left\langle \nabla f(x) - \widetilde{\nabla} f(x), y-x\right\rangle\right| \leqslant \Delta\|y-x\|$, т. е. для всяких $x, y \in Q$

$$f(y) \leqslant f(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle + \frac{L}{2}\|y-x\|^2 + \Delta\|y-x\|,$$

а также $f(y) \geqslant f(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle - \Delta\|y-x\|$.

Если кроме этого учесть неравенство $f_\delta(x) \leqslant f(x) \leqslant f_\delta(x) + \delta$ при $\delta > 0$, то получим следующий аналог (1.3):

$$f_\delta(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle - \Delta\|y-x\| \leqslant f(y)$$

$$\leqslant f_\delta(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle + \frac{L}{2}\|y-x\|^2 + \Delta\|y-x\| + \delta \ \ \forall x, y \in Q, \tag{1.6}$$

откуда $f_\delta(x) \leqslant f(x) \leqslant f_\delta(x) + \delta$.

В разд. 2 настоящей работы мы рассмотрим (в модельной общности подобно определению 1) следующий аналог неравенства (1.6) с параметрами $\delta, \gamma, \Delta \geqslant 0$:

$$f_\delta(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle - \gamma\|y-x\| \leqslant f(y)$$

$$\leqslant f_\delta(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle + \frac{L}{2}\|y-x\|^2 + \Delta\|y-x\| + \delta \ \ \forall x, y \in Q. \tag{1.7}$$

Смысл такого обобщения заключается в возможности различных значений параметров γ и Δ в (1.7). Один из основных результатов работы — обоснование потенциального уменьшения влияния Δ на оценку скорости сходимости метода. Отметим еще, что далее в разд. 3 подробно разобрано несколько примеров негладких задач, когда $\delta = \gamma = 0$ при $\Delta > 0$. Если положить $\gamma = 0$, то $\widetilde{\nabla} f(x)$ — δ-субградиент f в точке x, и параметр $\Delta > 0$ может указывать в этом случае на скачки $\widetilde{\nabla} f(x)$ в точках негладкости f. Если положить $\delta = 0$, то при $\gamma > 0$ $\widetilde{\nabla} f(x)$ — так называемый аналитический γ-субградиент f [10, Sect. 1.3]. В итоге мы предлагаем максимально общую концепцию неточной модели целевой функции, которая могла бы охватить все указанные ситуации. Для функций, допускающих существование такой модели в любой запрошенной точке, мы предлагаем адаптивный градиентный метод (алгоритм 1) и доказываем теорему о скорости его сходимости (теорема 1).

Неравенства (1.6) и (1.7) аналогичны (1.3), но величины $\Delta\|y-x\|$ и $\gamma\|y-x\|$ уже зависят от выбора x и y. Заменить их обе в (1.7) на постоянные величины, вообще говоря, возможно только в случае ограниченного допустимого множества задачи Q. Более того, хорошо известно, что при использовании $\widetilde{\nabla} f(x)$ из (1.5) метод может расходиться [6, Sect. 4]. Поэтому важно выделить класс задач, для которых можно получать приемлемые оценки скорости сходимости на неограниченных множествах. Это, в частности, мотивировало вторую часть основных результатов работы (разд. 4). Хорошо известно, что в случае сильной выпуклости целевого функционала оценки скорости сходимости градиентного метода могут существенно улучшаться. Например, для сильно выпуклого целевого функционала с липшицевым градиентом известно, что градиентный метод сходится с линейной скоростью. Весьма интересен и важен вопрос о том, насколько можно условие сильной выпуклости ослабить. В этом случае известен подход, основанный на использовании вместо сильной выпуклости условия градиентного доминирования Поляка — Лоясиевича [11] (см. также недавнюю работу [9] и имеющиеся там ссылки)

$$f(x)-f(x_*)\leqslant\frac{1}{2\mu}\|\nabla f(x)\|^2\quad\forall x\in Q, \tag{1.8}$$

где x_* — точное решение задачи минимизации f, а $\mu>0$ — некоторая постоянная. Известно, что неравенство (1.8) в предположении липшицевости градиента с константой L позволяет получить оценку скорости сходимости градиентного метода с постоянным шагом

$$f(x^N)-f(x_*)\leqslant\Big(1-\frac{\mu}{L}\Big)^N\big(f(x^0)-f(x_*)\big)\leqslant\exp\Big(-\frac{\mu}{L}N\Big)\big(f(x^0)-f(x_*)\big). \tag{1.9}$$

В настоящей работе мы рассматриваем следующий ослабленный вариант условия L-липшицевости градиента

$$f(y)\leqslant f(x)+\big\langle\widetilde{\nabla} f(x),y-x\big\rangle+\frac{L}{2}\|y-x\|^2+\Delta\|y-x\|+\delta\quad\forall\,x,y\in Q$$

для некоторых δ и $\Delta>0$. Например, это предположение естественно в случае, если значения функции f немного отличаются от значений некоторой достаточно гладкой функции $\widetilde{f}$, удовлетворяющей условию Липшица градиента (при этом $\widetilde{\nabla} f(x)$ — некоторое возмущенное с точностью Δ значение градиента $\nabla f(x)$). По сути, в разд. 4 работы левая часть неравенства (1.7) заменяется условием градиентного доминирования. Мы предлагаем метод с адаптивным подбором шага с настройкой на величины L, Δ и δ и показываем оценку скорости сходимости, аналогичную (1.9). В частности, запуск предлагаемого метода (алгоритм 2) не предполагает знания никакой верхней оценки L и может применяться для задач с неточным заданием градиента на неограниченных допустимых множествах. Более того, возможно использование данного подхода для некоторого класса негладких задач (см. определение 3).

Подытожим основные результаты (вклад) настоящей работы:

— В раз. 2 обобщено ранее предложенное в [2] понятие (δ,L)-модели целевой функции в запрошенной точке и введена концепция (δ,γ,Δ,L)-модели функции (определение 2). Предложен градиентный метод (алгоритм 1) для задач выпуклой минимизации с адаптивным выбором шага и адаптивной настройкой на некоторые из параметров (δ,γ,Δ,L)-модели, получена оценка качества решения в зависимости от номера итерации (теорема 1).

— В разд. 3 рассмотрен специальный класс задач выпуклой негладкой оптимизации, к которым применима концепция определения 2 ($\delta=\gamma=0$, $\Delta>0$). Показано, что для таких задач возможно модифицировать алгоритм 1 так, чтобы гарантированно имела место сходимость по функции со скоростью $O(\varepsilon^{-2}\log_2\varepsilon^{-1})$, которая близка к оптимальной на классе задач выпуклой негладкой оптимизации (теорема 2). При этом рассмотрены примеры негладких задач, для которых за счет адаптивности алгоритма 1 может наблюдаться существенно более высокая скорость сходимости.

— В разд. 4 предложен адаптивный градиентный метод (алгоритм 2) для целевых функционалов с липшицевым градиентом (а также некоторой релаксацией этого условия), удовлетворяющих условию Поляка — Лоясиевича. При этом учитывается возможность неточного задания градиента и предлагается адаптивная настройка работы метода на основные входные параметры. Обоснована линейная сходимость метода с точностью до величины, связанной с погрешностью (теорема 3).

2. Концепция $(\delta, \gamma, \Delta, L)$-модели функции в запрошенной точке и оценка скорости сходимости для градиентного метода

Введем анонсированный выше аналог понятия $(\delta, \gamma, \Delta, L)$-модели целевой функции, который учитывает погрешность Δ задания градиента и применим также для задач с относительно гладкими целевыми функционалами [4].

О п р е д е л е н и е 2. Будем говорить, что f допускает $(\delta, \gamma, \Delta, L)$-модель в точке $x \in Q$, если для некоторой выпуклой по первой переменной функции $\psi(y, x)$ такой, что $\psi(x, x) = 0$ для произвольных $x, y \in Q$, будет верно неравенство

$$f_\delta(x) + \psi(y, x) - \gamma\|y - x\| \leqslant f(y) \leqslant f_\delta(x) + \psi(y, x) + \delta + \Delta\|y - x\| + LV(y, x). \tag{2.1}$$

Покажем пример, поясняющий смысл использования модельной общности в предыдущем определении.

П р и м е р 1. Отметим задачу выпуклой композитной оптимизации $f(x) = g(x) + h(x) \to \min$, где g — гладкая выпуклая функция, а h — выпуклая необязательно гладкая функция простой структуры (операция проектирования на любое множество уровня h несильно затратна). Если при этом для градиента ∇g задано его приближение $\widetilde{\nabla} g$: $\left\|\widetilde{\nabla} g(x) - \nabla g(x)\right\| \leqslant \Delta$, причем $g(y) \geqslant g(x) + \langle \widetilde{\nabla} g(x), y - x\rangle - \gamma\|y - x\| - \delta$, то можно положить $\psi(y, x) = \langle \widetilde{\nabla} g(x), y - x\rangle + h(y) - h(x)$, и будет верно (2.1). Композитная оптимизация весьма часто возникает во многих прикладных задачах (см., например, [7]).

Рассмотрим следующий метод для минимизации выпуклых функций, которые допускают существование $(\delta, \gamma, \Delta, L)$-модели во всякой точке $x \in Q$ и докажем результат о его скорости сходимости.

А л г о р и т м 1: Адаптивный градиентный метод для выпуклых функций, допускающих $(\delta, \gamma, \Delta, L)$-модель в произвольной запрошенной точке.

Require: x^0 — начальная точка, $V(x_*, x^0) \leqslant R^2$, параметры $\delta_0,\ L_0,\ \Delta_0$ $(\delta_0 \leqslant 2\delta,\ L_0 \leqslant 2L,\ \Delta_0 \leqslant 2\Delta)$.

1: $L_{k+1} := {}^{L_k}/_2,\ \Delta_{k+1} := {}^{\Delta_k}/_2,\ \delta_{k+1} := {}^{\delta_k}/_2$.
2: $x^{k+1} := \arg\min\limits_{x \in Q}\{\psi(x, x^k) + LV(x, x^k)\}$.
3: **if** $f_\delta(x^{k+1}) \leqslant f_\delta(x^k) + \psi(x^{k+1}, x^k) + L_{k+1}V(x^{k+1}, x^k) + \Delta_{k+1}\left\|x^{k+1} - x^k\right\| + \delta_{k+1}$ **then**
4: $\quad k := k + 1$ и выполнение п. 1.
5: **else**
6: $\quad L_{k+1} := 2 \cdot L_{k+1};\ \Delta_{k+1} := 2 \cdot \Delta_{k+1};\ \delta_{k+1} := 2 \cdot \delta_{k+1}$ и выполнение п. 2.
7: **end if**

Ensure: $\widehat{x} := \frac{1}{S_N}\sum\limits_{k=0}^{N-1}\frac{x^{k+1}}{L_{k+1}},\ S_N := \sum\limits_{k=0}^{N-1}\frac{1}{L_{k+1}}$.

Справедливо следующее утверждение.

Теорема 1. *Пусть $f : Q \to \mathbb{R}$ — выпуклая функция и для некоторой постоянной $R > 0$ имеет место $V(x_*, x^0) \leqslant R^2$, где x^0 — начальное приближение, а x_* — точное решение*

задачи минимизации f, ближайшее к x^0 с точки зрения расхождения Брэгмана. Тогда после N итераций для выхода $\hat{x}$ алгоритма 1 *будет верно неравенство*

$$f(\hat{x}) - f(x_*) \leqslant \frac{R^2}{S_N} + \frac{2}{S_N}\sum_{k=0}^{N-1}\frac{\delta_{k+1} + \Delta_{k+1}\|x^{k+1} - x^k\| + \gamma\|x^k - x_*\|}{L_{k+1}} + \delta. \tag{2.2}$$

При этом количество обращений к задаче п. 2 *листинга алгоритма* 1 *не превышает*

$$2N + \max\Big\{\log_2\frac{2L}{L_0}, \log_2\frac{2\delta}{\delta_0}, \log_2\frac{2\Delta}{\Delta_0}\Big\}. \tag{2.3}$$

Доказательство. 1. Согласно лемме 1 из [2] после завершения k-й итерации ($k = 0, 1, 2, \ldots$) алгоритма 1 будут верны неравенства

$$\psi(x^{k+1}, x^k) \leqslant \psi(x, x^k) + L_{k+1}V(x, x^k) - L_{k+1}V(x, x^{k+1}) - L_{k+1}V(x^{k+1}, x^k),$$

$$f_\delta(x^{k+1}) \leqslant f_\delta(x^k) + \psi(x^{k+1}, x^k) + L_{k+1}V(x, x^k) - L_{k+1}V(x, x^{k+1}) + \Delta_{k+1}\|x^{k+1} - x^k\| + \delta_{k+1}.$$

Поэтому ввиду того, что $f_\delta(x) \leq f(x) \leq f_\delta(x) + \delta$ при всяком $x \in Q$ имеем

$$f(x^{k+1}) \leqslant f(x^k) + \psi(x, x^k) + L_{k+1}V(x, x^k) - L_{k+1}V(x, x^{k+1}) + \Delta_{k+1}\|x^{k+1} - x^k\| + \delta_{k+1} + \delta.$$

Далее, с учетом левой части неравенства (2.1) при $x = x_*$ получим

$$f(x^{k+1}) - f(x_*) \leqslant L_{k+1}V(x_*, x^k) - L_{k+1}V(x_*, x^{k+1}) + \Delta_{k+1}\|x^{k+1} - x^k\| + \delta_{k+1} + \delta + \gamma\|x^k - x_*\|,$$

откуда после суммирования по $k = 0, 1, \ldots, N-1$ ввиду выпуклости f имеем

$$f(\hat{x}) - f(x_*) \leqslant \frac{1}{S_N}\sum_{k=0}^{N-1}\frac{f(x^{k+1})}{L_{k+1}} - f(x_*) \leqslant V(x_*, x^0)$$

$$+ \frac{1}{S_N}\sum_{k=0}^{N-1} L_{k+1}^{-1}\Big(\Delta_{k+1}\|x^{k+1} - x^k\| + \delta_{k+1} + \gamma\|x^k - x_*\|\Big) + \delta.$$

2. Проверим оценку (2.3). Пусть на $(k+1)$-й итерации ($k = 0, 1, \ldots, N-1$) алгоритма 1 вспомогательная задача решается i_{k+1} раз. Тогда

$$2^{i_{k+1}-2} = \frac{L_{k+1}}{L_k} = \frac{\delta_{k+1}}{\delta_k} = \frac{\Delta_{k+1}}{\Delta_k},$$

поскольку в начале каждой итерации параметры L_k, δ_k, Δ_k делятся на 2. Поэтому

$$\sum_{k=0}^{N-1} i_{k+1} = 2N + \log_2\frac{L_N}{L_0}, \quad \log_2\frac{L_N}{L_0} = \log_2\frac{\delta_N}{\delta_0} = \log_2\frac{\Delta_N}{\Delta_0}.$$

Ясно, что верно хотя бы одно из неравенств $L_N \leqslant 2L$, $\delta_N \leqslant 2\delta$ и $\Delta_N \leqslant 2\Delta$, что и обосновывает оценку (2.3). □

Замечание 1. Оценка (2.3) показывает, что в среднем трудоемкость итерации предложенного адаптивного алгоритма превышает трудоемкость аналогичного неадаптивного метода с постоянным шагом не более, чем в постоянное число раз. Отметим также, что при $k = 0, 1, 2, \ldots$ верно $L_{k+1} \leqslant 2CL$, где $C = \max\Big\{1, \frac{2\delta}{\delta_0}, \frac{2\Delta}{\Delta_0}\Big\}$. Поэтому $S_N \leqslant \dfrac{N}{2CL}$, что указывает на скорость сходимости метода $O(\varepsilon^{-1})$, но при наличии в оценке (2.2) слагаемых, определяемых параметрами δ, γ, Δ (при этом ввиду адаптивности метода δ_k и Δ_k могут быть меньше δ и Δ соответственно). Можно доказать, что эта величина ограничена в случае ограниченного допустимого множества задачи Q, что вполне может считаться оптимальным [12].

З а м е ч а н и е 2. Если дополнительно предположить, что в произвольной точке $x \in Q$ верно $f_\delta(x) = f(x)$, то в оценке (2.2) можно считать $\delta = 0$. В таком случае оценка (2.2) полностью адаптивна по параметрам L, Δ и δ.

З а м е ч а н и е 3. Отметим, что ввиду адаптивности алгоритма 1 полученная в теореме 1 оценка скорости сходимости может быть применена даже в случаях $L = +\infty$ или $\Delta = +\infty$. Если не происходит зацикливания и каждый раз выполняется критерий выхода из итерации, то алгоритм 1 применим и в этом случае. Пример, когда такое возможно ($\Delta = +\infty$), приведен в следующем разделе (задача 2).

3. О применимости метода к одному классу негладких задач за счет введения искусственных неточностей

Отметим, что на величину Δ в (1.5) можно смотреть как на искусственную неточность, описывающую степень негладкости функционала f. Точнее говоря, Δ можно понимать, например, как верхнюю оценку суммы диаметров субдифференциалов f в точках негладкости вдоль всевозможных векторных отрезков $[x; y]$ из области определения f. В [13] введен следующий класс негладких выпуклых функционалов.

О п р е д е л е н и е 3. Будем говорить, что выпуклый функционал $f\colon Q \to \mathbb{R}$ ($Q \subset \mathbb{R}^n$) имеет (δ, L)-липшицев субградиент ($f \in C^{1,1}_{L,\hat{\Delta}}(Q)$), если для некоторых $\delta > 0$ и $L > 0$

(i) для произвольных $x, y \in Q$ выпуклый функционал f дифференцируем во всех точках множества $\{y_t\}_{0\leq t\leq 1}$ ($y_t = (1-t)x + ty$), за исключением последовательности (возможно, конечной)

$$\{y_{t_j}\}_{j=1}^{\infty} :\ t_1 < t_2 < t_3 < \ldots \text{ и } \lim_{j\to\infty} t_j = 1; \tag{3.1}$$

(ii) для последовательности точек из (3.1) существуют конечные субдифференциалы в смысле выпуклого анализа $\{\partial f(y_{t_j})\}_{j=1}^{\infty}$ и

$$\operatorname{diam} \partial f(y_{t_j}) =: \hat{\Delta}_j > 0, \quad \text{где } \sum_{j=1}^{+\infty} \hat{\Delta}_j =: \hat{\Delta} < +\infty;$$

(iii) для произвольных $x, y \in Q$ при условии, что $y_t \in Q \setminus Q_0$ при всяком $t \in (0,1)$ (то есть существует градиент $\nabla f(y_k)$) для некоторой фиксированной константы $L > 0$, не зависящей от выбора x и y, выполняется неравенство

$$\min_{\nabla f(x)\in\partial f(x),\ \nabla f(y)\in\partial f(y)} \|\nabla f(x) - \nabla f(y)\|_* \leqslant L\|x - y\|.$$

В [13], в частности, показано, что всякий функционал $f \in C^{1,1}_{L,\delta}(Q)$ удовлетворяет для произвольного субградиента $\nabla f(x) \in \partial f(x)$ неравенству

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 + 2\hat{\Delta}\|y - x\| \quad \forall y \in Q. \tag{3.2}$$

С другой стороны, ввиду выпуклости f будет верно $f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle$. Поэтому всякая функция $f \in C^{1,1}_{L,\hat{\Delta}}(Q)$ удовлетворяет определению 2 при $\psi(y, x) = \langle \nabla f(x), y - x\rangle$ с параметрами $\delta = \gamma = 0$ и $\Delta = 2\hat{\Delta}$.

Оказывается, что экспериментально можно получать существенно лучшую скорость сходимости метода, чем по отмеченной выше оценке. Приведем некоторые примеры. Начнем с примера задачи (см., например, [14]) с бесконечным числом точек негладкости (недифференцируемости) целевого функционала, но с конечной величиной Δ.

Результаты численных экспериментов

	Задача 1					Задача 2				
Итерации	200	400	600	800	1000	200	400	600	800	1000
Оценка	0.0232	0.0117	0.0079	0.006	0.0048	0.79	0.44	0.31	0.24	0.2
Время, с	27	54	82	110	136	15	29	44	58	72

З а д а ч а 1 (Аналог задачи Ферма — Торричелли — Штейнера). Для заданных шаров ω_k с центрами $a_k = (a_{1k}, a_{2k}, \ldots, a_{nk})$ (координаты точек a_k выбираются случайно так, чтобы $1 < \sqrt{a_{1k}^2 + a_{2k}^2 + \ldots + a_{nk}^2} < 1.5$, $k = \overline{1,m}$, $m = 10$) и единичными радиусами в n-мерном евклидовом пространстве $\mathbb{R}^n$ ($n = 10^5$) необходимо найти такую точку $x = (x_1, x_2, \ldots, x_n)$, чтобы целевая функция $f(x) := \sum_{k=1}^{m} d(x, \omega_k)$ принимала наименьшее значение на множестве точек единичного шара с центром в нуле, где $d(x, \omega_k) = \|x - a_k\| - 1$, если $\|x - a_k\| > 1$, и 0 в противном случае (здесь $\|x - a_k\| = \sqrt{(x_1 - a_{1k})^2 + (x_2 - a_{2k})^2 + \ldots + (x_n - a_{nk})^2}$).

В таблице выше для задачи 1 приведены усредненные результаты 10 экспериментов со случайным подбором координат точек для указанного количества итераций. Как видим, скорость сходимости метода близка к $O(\varepsilon^{-1})$. Это свойственно для неускоренных градиентных методов на классе задач оптимизации выпуклых функций с липшицевым градиентом (так называемых гладких задач). Однако рассматриваемая задача негладкая, поскольку точки недифференцируемости f лежат в области определения (в единичном шаре с центром в нуле). Для задач минимизации выпуклых липшицевых функций, как известно, оптимальная оценка скорости сходимости (суб)градиентных методов — $O(\varepsilon^{-2})$ [15]. Оценку $O(\varepsilon^{-1})$ можно объяснить адаптивностью предложенного метода.

Рассмотрим еще пример, где довольно много точек негладкости. В частности, все точки некоторого векторного отрезка могут быть точками негладкости, и условие (3.2) выполнено лишь для бесконечного значения Δ.

З а д а ч а 2 (Задача о наименьшем покрывающем шаре). Для заданных точек

$$a_k = (a_{1k}, a_{2k}, \ldots, a_{nk})$$

найти евклидов шар наименьшего радиуса, в котором лежат эти точки. Координаты точек a_k выбираются случайно так, что $0.5 < \sqrt{a_{1k}^2 + a_{2k}^2 + \ldots + a_{nk}^2} < 1$, $k = \overline{1,10}$, в n-мерном евклидовом пространстве $\mathbb{R}^n$ (размерность $n = 10^5$) необходимо найти такую точку $x = (x_1, x_2, \ldots, x_n)$, чтобы целевая функция

$$f(x) := \max_{k=\overline{1,m}} \|x - a_k\| = \max_{k=\overline{1,m}} \sqrt{(x_1 - a_{1k})^2 + (x_2 - a_{2k})^2 + \ldots + (x_n - a_{nk})^2}$$

принимала наименьшее значение. Мы рассматриваем задачу нахождения наиболее подходящей точки на единичном шаре с центром в нуле.

В таблице выше для задачи 2 приведены усредненные результаты 10 экспериментов со случайным подбором координат точек для определенного количества итераций. Как видим, скорость сходимости метода снова близка к $O(\varepsilon^{-1})$.

Приведенные результаты экспериментов указывают на потенциально неплохую эффективность предложенной адаптивной процедуры регулировки шага в методе. Отметим, что все вычисления были произведены с помощью программного обеспечения CPython 3.7 на компьютере с 3-ядерным процессором AMD Athlon II X3 450 с тактовой частотой 3,2 ГГц на каждое ядро. ОЗУ компьютера составляло 8 Гб.

Однако можно в некотором смысле и теоретически показать оптимальность предложенной схемы для рассматриваемых негладких задач. Оказывается, в случае известной величины

$\Delta < +\infty$ возможно несколько модифицировать алгоритм 1, обеспечив уменьшение $\Delta_k\|x^{k+1} - x^k\|$ в (2.2) до любой заданной величины. Это позволит показать оптимальность данного метода в теории нижних оракульных оценок [15] с точностью до логарифмического множителя.

Покажем, как это возможно сделать. Предположим, что на (k+1)-й итерации алгоритма 1 ($k = 0, 1, \ldots, N-1$) верно неравенство $L \leqslant L_{k+1} \leqslant 2L$ (как показано в п. 2 доказательства теоремы 1, этого можно всегда добиться выполнением не более чем постоянного числа операций п. 2 листинга алгоритма 1). Для каждой итерации алгоритма 1 ($k = 0, 1, \ldots, N-1$) предложим такую процедуру

Повторяем операции п. 2 p раз, увеличивая L_{k+1} в два раза при неизменной $\Delta_{k+1} \leqslant 2\Delta$.	(3.3)

Процедуру (3.3) остановим в случае выполнения одного из неравенств

$$\Delta_{k+1}\left\|x^{k+1} - x^k\right\| \leqslant \frac{\varepsilon}{2} \tag{3.4}$$

или

$$f(x^{k+1}) \leqslant f(x^k) + \left\langle \widetilde{\nabla} f(x^k), x^{k+1} - x^k \right\rangle + 2^{p-1}L\left\|x^{k+1} - x^k\right\|^2. \tag{3.5}$$

Отметим, что здесь мы полагаем f точно заданной, т. е. $f_\delta = f$ ($\delta = 0$); $\widetilde{\nabla} f$ — некоторый субградиент f. Процедура (3.3) предполагает на $(k+1)$-й итерации ($k = 0, 1, 2, \ldots, N-1$) обновления x^{k+1} (при сохранении x^k). Оценим количество повторений p шага п. 2 листинга алгоритма 1, необходимое для достижения альтернативы (3.4), (3.5). Для всяких $x^k, x^{k+1} \in Q$ по предположению верно неравенство

$$f(x^{k+1}) \leqslant f(x^k) + \left\langle \widetilde{\nabla} f(x^k), x^{k+1} - x^k \right\rangle + \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 + \Delta\left\|x^{k+1} - x^k\right\|,$$

причем $\Delta_{k+1} \leqslant 2\Delta$. Если не выполнено (3.4), то $\left\|x^{k+1} - x^k\right\| > \dfrac{\varepsilon}{4\Delta}$ и (3.5) заведомо верно при

$$2^p > 1 + \frac{16\Delta^2}{\varepsilon L}, \tag{3.6}$$

поскольку в таком случае

$$\frac{2^p - 1}{2}L\left\|x^{k+1} - x^k\right\|^2 > 2\Delta\left\|x^{k+1} - x^k\right\|.$$

Итак, после повторения p процедур (p удовлетворяет (3.6)) типа (3.3) на каждой из N итераций алгоритма 1 неравенство (2.2) примет вид

$$f(\hat{x}) - f^* \leqslant \frac{R^2}{S_N} + \frac{\varepsilon}{2}, \quad \text{где} \quad S_N = \sum_{k=0}^{N-1} \frac{1}{L_{k+1}} \geqslant \frac{N}{2^{p+1}L}.$$

Поэтому $\dfrac{R^2}{S_N} \leqslant \dfrac{2^{p+1}LR^2}{N} \leqslant \dfrac{\varepsilon}{2}$ в случае $N \geqslant \dfrac{2^{p+2}LR^2}{\varepsilon}$. С учетом (3.6) получаем оценку

$$N \geqslant \frac{4LR^2}{\varepsilon} + \frac{64\Delta^2R^2}{\varepsilon^2}.$$

При этом (3.6) означает, что на каждой итерации потребуется не более чем

$$p = \left\lceil \log_2\left(1 + \frac{16\Delta^2}{\varepsilon L}\right)\right\rceil$$

шагов типа п. 2 листинга алгоритма 1 (т. е. операций проектирования) для стандартной модели $\psi(y, x) = \langle \nabla f(x), y - x\rangle$. Итак, верна

Теорема 2. *Пусть функция f удовлетворяет определению 2 при $\psi(y,x)=\langle\nabla f(x),y-x\rangle$ с параметрами $\delta=\gamma=0$ и $\Delta>0$. Тогда в обозначениях теоремы 1 для выхода $\hat{x}$ модифицированного алгоритма 1 с учетом дополнительной процедуры (3.3) неравенство $f(\hat{x})-f^*\leqslant\varepsilon$ будет гарантированно выполнено не более чем после*

$$\left\lceil\left(\frac{4LR^2}{\varepsilon}+\frac{64\Delta^2R^2}{\varepsilon^2}\right)\right\rceil\cdot\left\lceil\log_2\left(1+\frac{16\Delta^2}{\varepsilon L}\right)\right\rceil \tag{3.7}$$

вычислений субградиента f.

Таким образом доказано, что для рассмотренного класса негладких задач приемлемое качество решения можно достичь за $O\big(\varepsilon^{-2}\log_2\varepsilon^{-1}\big)$ вычислений субградиента f, что близко к оптимальной оценке с точностью до логарифмического множителя. Отметим, что примеры сходимости метода со скоростью $O(\varepsilon^{-1})$ для некоторых выпуклых негладких задач наблюдались и для так называемого универсального градиентного метода [16] с другой концепцией искусственной неточности. Однако для негладких задач с липшицевым целевым функционалом в [16] доказана оценка скорости сходимости вида $O\big(M_f\varepsilon^{-2}\big)$, зависящая еще от константы Липшица целевого функционала M_f. Полученная нами оценка (3.7) может быть лучше при малом $\Delta>0$ (в этом случае оценка (3.7) близка к $O\big(\varepsilon^{-1}\big)$).

4. Метод для минимизации функций, удовлетворяющих условию градиентного доминирования при неточном задании целевой функции и градиента

Теперь предложим подход к задаче минимизации, вообще говоря, невыпуклых функций с неточно заданным градиентом. При этом метод предполагает адаптивную настройку на некоторые параметры, в том числе связанные с величиной погрешности задания градиента. Пусть рассматривается задача минимизации функции на всем пространстве $f:\mathbb{R}^n\to\mathbb{R}$, для которой

(i) существует $x_*\in\mathbb{R}^n$ такое, что

$$f(x_*)=\min_{x\in\mathbb{R}^n}f(x)=:f^*; \tag{4.1}$$

(ii) выполнено условие Поляка — Лоясиевича (1.8) (или (PL)-условие);

(iii) для некоторых постоянных $L>0$ и $\Delta>0$ верно неравенство

$$f(y)\leqslant f(x)+\langle\nabla f(x),y-x\rangle+\frac{L\|y-x\|^2}{2}\quad\forall x,y\in\mathbb{R}^n, \tag{4.2}$$

где норма $\|\cdot\|$ евклидова.

Если предположить, что в каждой точке $x\in\mathbb{R}^n$ доступно приближенное значение $\widetilde{\nabla}f(x)$ градиента $\nabla f(x)$: $\big\|\widetilde{\nabla}f(x)-\nabla f(x)\big\|\leqslant\Delta\quad\forall x\in\mathbb{R}^n$ при некотором фиксированном $\Delta>0$, то (4.2) верно с заменой градиента $\nabla f(x)$ на $\widetilde{\nabla}f(x)$. Далее для удобства будем обозначать $g_x:=\|\nabla f(x)\|$ и $\widetilde{g}_x:=\|\widetilde{\nabla}f(x)\|$. К задаче (4.1) будем применять градиентный метод вида

$$x^{k+1}=x^k-h_k\widetilde{\nabla}f(x^k), \tag{4.3}$$

$k=0,1,2,\ldots$ и $h_k>0$. При этом h_k выберем так, чтобы

$$f(x^{k+1})\leqslant f(x^k)+\big\langle\widetilde{\nabla}f(x^k),x^{k+1}-x^k\big\rangle+\frac{L\big\|x^{k+1}-x^k\big\|^2}{2}+\Delta_{k+1}\big\|x^{k+1}-x^k\big\|, \tag{4.4}$$

где $\Delta_{k+1} > 0$ — адаптивно подбираемая величина. В начале каждой итерации $\Delta_{k+1} := \dfrac{\Delta_k}{2}$, а далее Δ_{k+1} $(k = 0, 1, 2, \ldots)$ увеличивается в два раза, и процедура (4.3) повторяется до тех пор, пока не выполняется (4.4).

Ясно, что (4.4) заведомо верно при $\Delta_{k+1} \geqslant \Delta$. Поэтому аналогично п. 2) доказательства теоремы 1 проверяется, что за конечное число таких шагов (4.4) будет выполнено на любой итерации $(k = 0, 1, 2, \ldots)$, после чего

$$f(x^{k+1}) - f(x^k) \leqslant \varphi(h_k), \text{ где } \varphi(h) = -h\widetilde{g}^2_{x^k} + \frac{Lh^2}{2}\widetilde{g}_{x^k} + 2h\widetilde{g}_{x^k}.$$

Выберем шаг h_k так, чтобы минимизировать величину $\varphi(h_k)$, т. е. $\varphi'(h_k) = 0$, и

$$h_k = \frac{1}{L} - \frac{\Delta_{k+1}}{L\widetilde{g}_{x^k}}.$$

В таком случае (4.4) означает, что

$$f(x^{k+1}) - f(x^k) \leqslant -\frac{1}{2L}\big(\widetilde{g}_{x^k} - \Delta_{k+1}\big)^2 \leqslant -\frac{1}{2L}\Big(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}\Big)g^2_{x^k}, \tag{4.5}$$

поскольку $|\widetilde{g}_{x^k} - g_{x^k}| \leqslant \big\|\widetilde{\nabla} f(x^k) - \nabla f(x^k)\big\| \leqslant \Delta$ и $\widetilde{g}_{x^k} + \Delta \geqslant g_{x^k}$. Неравенство (4.5) означает, что для произвольного $k = 0, 1, 2, \ldots$

$$f(x^k) - f(x^{k+1}) \geqslant \frac{1}{2L}\Big(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}\Big)^2 g^2_{x^k} \geqslant \frac{\mu}{L}\Big(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}\Big)^2 \big(f(x^k) - f^*\big)$$

ввиду (PL)-условия (1.8). Поэтому

$$f(x^{k+1}) - f(x_*) \leqslant \Big(1 - \frac{\mu}{L}\Big(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}\Big)^2\Big)\big(f(x^k) - f(x_*)\big),$$

откуда

$$f(x^{k+1}) - f^* \leqslant \prod_{i=0}^{k}\Big(1 - \frac{\mu}{L}\Big(\frac{\widetilde{g}_{x^i} - \Delta_{i+1}}{\widetilde{g}_{x^i} + \Delta}\Big)^2\Big)\big(f(x^0) - f^*\big). \tag{4.6}$$

Можно считать, что $\mu \leqslant L$, и ввиду $\widetilde{g}_{x^i} - \Delta_{i+1} < \widetilde{g}_{x^i} + \Delta$ в (4.6) справа входит произведение $k + 1$ числа, каждое из которых меньше 1. Адаптивность подбора $\Delta_{k+1} \leqslant 2\Delta$ на каждой итерации может привести к увеличению дроби $\dfrac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}$ и уменьшению множителей в (4.6), что потенциально улучшает оценку по сравнению с неадаптивным вариантом

$$f(x^{k+1}) - f^* \leqslant \prod_{i=0}^{k}\Big(1 - \frac{\mu}{L}\Big(\frac{\widetilde{g}_{x^i} - \Delta}{\widetilde{g}_{x^i} + \Delta}\Big)^2\Big)\big(f(x^0) - f^*\big). \tag{4.7}$$

Аналогичную оценку можно предложить, также используя следующий метод с адаптивным выбором не только погрешности на итерациях, но и величины L (см. алгоритм 2).

Для данного алгоритма будет верна оценка (4.8), обоснование которой аналогично (4.7). Более того, можно показать, что либо невязка $\min\limits_k f(x^k) - f^*$ убывает со скоростью геометрической прогрессии при увеличении k (см. (4.9)), либо она ограничена величиной Δ (см. (4.10)). Справедливо следующее утверждение.

А л г о р и т м 2: Адаптивный градиентный метод для функций, удовлетворяющих (PL)-условию.

Require: x^0 — начальная точка, параметры $\Delta_0,\ L_0$
$(2\mu \leqslant L_0 < 2L,\ \Delta_0 \leqslant 2\Delta)$.
1: $L_{k+1} := {}^{L_k}/_2,\ \Delta_{k+1} := {}^{\Delta_k}/_2$.
2: $x^{k+1} = x^k - h_k \widetilde{\nabla} f(x^k)$,

$$h_k = \frac{1}{L_{k+1}} - \frac{\Delta_{k+1}}{L_{k+1}\widetilde{g}_{x^k}},\ \widetilde{g}_{x^k} = \|\widetilde{\nabla} f(x^k)\|.$$

3: **repeat**
4: **if** $f(x^{k+1}) \leqslant f(x^k) + \left\langle \widetilde{\nabla} f(x^k), x^{k+1} - x^k \right\rangle + \frac{L_{k+1}}{2}\|x^{k+1} - x^k\|^2 + \Delta_{k+1}\|x^{k+1} - x^k\|$ **then**
5: $k := k + 1$ и выполнение п. 1.
6: **else**
7: $L_{k+1} := 2 \cdot L_{k+1};\ \Delta_{k+1} := 2 \cdot \Delta_{k+1}$ и выполнение п. 2.
8: **end if**
9: **until** $k >= N$
Ensure: x^{k+1}.

Теорема 3. *После k итераций алгоритма 2 будет выполняться следующее неравенство:*

$$f(x^{k+1}) - f^* \leqslant \prod_{i=0}^{k}\Big(1 - \frac{\mu}{L_{k+1}}\Big(\frac{\widetilde{g}_{x^i} - \Delta_{i+1}}{\widetilde{g}_{x^i} + \Delta}\Big)^2\Big)\big(f(x^0) - f^*\big). \tag{4.8}$$

Более того, если дополнительно потребовать для алгоритма 2 $\Delta_{k+1} = \min\{\Delta_{k+1}, \Delta\}$, то для всякого $C > 1$ будет выполняться одно из двух неравенств

$$f(x^{k+1}) - f^* \leqslant \Big(1 - \frac{\mu}{L}\Big(\frac{C-1}{C+1}\Big)^2\Big)^{k+1}\big(f(x^0) - f^*\big) \tag{4.9}$$

или

$$\min_{i=\overline{1,k+1}} f(x^i) - f^* < \frac{(C+1)^2\Delta^2}{2\mu}. \tag{4.10}$$

Д о к а з а т е л ь с т в о. Отметим лишь, что при произвольном $k \geqslant 0$ верно

$$\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta} \geqslant \frac{\widetilde{g}_{x^k} - \Delta}{\widetilde{g}_{x^k} + \Delta} = 1 - \frac{2\Delta}{\widetilde{g}_{x^k} + \Delta}.$$

Пусть $\widetilde{g}_{x^k} \geqslant C\Delta$ для некоторой постоянной $C > 1$. Тогда $1 - \dfrac{2\Delta}{\widetilde{g}_{x^k} + \Delta} \geqslant 1 - \dfrac{2}{C+1} = \dfrac{C-1}{C+1} > 0$ и (4.6) принимает вид (4.9). Если же для некоторого k верно $\widetilde{g}_{x^k} < C\Delta$, то $g_{x^k} < C\Delta + \Delta = \Delta(C+1)$, и (4.10) верно в силу (PL)-условия (1.8). □

З а м е ч а н и е 4. Можно рассматривать вместо (4.3) более слабое условие

$$f_\delta(y) \leqslant f_\delta(x) + \langle \widetilde{\nabla} f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 + \Delta\|y - x\| + \delta \quad \forall\, x, y \in Q \tag{4.11}$$

для некоторого $\delta > 0$ и приближения f_δ: $f_\delta(x) \leq f(x) \leq f_\delta + \delta$. Например, это актуально в случае, если значения f немного отличаются от значений некоторой достаточно гладкой функции $\widetilde{f}$, удовлетворяющей (4.3) (при этом $\widetilde{\nabla} f(x)$ — некоторое возмущенное с точностью Δ значение градиента $\nabla\widetilde{f}(x)$). Тогда рассмотрим метод (4.4), (4.5) с видоизмененным критерием выхода из итерации

$$f_\delta(x^{k+1}) \leqslant f_\delta(x^k) + \langle \widetilde{\nabla} f(x^k), x^{k+1} - x^k\rangle + \frac{L_{k+1}\|x^{k+1} - x^k\|^2}{2} + \Delta_{k+1}\|x^{k+1} - x^k\| + \delta_{k+1},$$

который предполагает адаптивный подбор величин Δ_{k+1} и δ_{k+1} при заданных изначально $L_0 \leqslant 2L$, $\Delta_0 \leqslant \Delta$ и $\delta_0 \leqslant 2\delta$. Тогда на каждой итерации вместо неравенства (4.5) будет верно

$$f(x^k) - f(x^{k+1}) + \delta_{k+1} + \delta \geqslant f_\delta(x^k) - f_\delta(x^{k+1}) + \delta_{k+1} \geqslant \frac{\mu}{L_{k+1}}\Big(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}\Big)^2\big(f(x^k) - f^*\big),$$

откуда аналогично (4.6) имеем

$$f(x^{k+1}) - f^* \leqslant \Big(1 - \frac{\mu}{L_{k+1}}\Big(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}\Big)^2\Big)\big(f(x^k) - f^*\big) + \delta_{k+1} + \delta$$

$$\leqslant \Big(1 - \frac{\mu}{L_{k+1}}\Big(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}\Big)^2\Big)\Big(1 - \frac{\mu}{L_k}\Big(\frac{\widetilde{g}_{x^{k-1}} - \Delta_k}{\widetilde{g}_{x^{k-1}} + \Delta}\Big)^2\Big)\big(f(x^{k-1}) - f^*\big)$$

$$+ \ (\delta_k + \delta)\Big(1 - \frac{\mu}{L_{k+1}}\Big(\frac{\widetilde{g}_{x^{k-1}} - \Delta_k}{\widetilde{g}_{x^{k-1}} + \Delta}\Big)^2\Big) + \delta + \delta_{k+1} \leqslant \ldots$$

$$\leqslant \prod_{i=0}^{k}\Big(1 - \frac{\mu}{L_{i+1}}\Big(\frac{\widetilde{g}_{x^i} - \Delta_{i+1}}{\widetilde{g}_{x^i} + \Delta}\Big)^2\Big)\big(f(x^0) - f^*\big) + \sum_{i=0}^{k-1}(\delta + \delta_{i+1})\prod_{j=i}^{k}\Big(1 - \frac{\mu}{L_{j+1}}\Big(\frac{\widetilde{g}_{x^j} - \Delta_{j+1}}{\widetilde{g}_{x^j} + \Delta}\Big)^2\Big) + \delta_{k+1} + \delta.$$

Полученная оценка выглядит несколько громоздко. Конкретизируя ее при постоянном $L_{i+1} = L$, $\Delta = 0$ и $\delta_{i+1} \leqslant 2\delta$ $(i \geqslant 0)$, получаем

$$f(x^{k+1}) - f^* \leqslant \Big(1 - \frac{\mu}{L}\Big)^{k+1}\big(f(x^0) - f^*\big) + \sum_{i=0}^{k}(\delta + \delta_{i+1})\Big(1 - \frac{\mu}{L}\Big)^{k-i}$$

$$\leqslant \Big(1 - \frac{\mu}{L}\Big)^{k+1}\big(f(x^0) - f^*\big) + 3\delta\sum_{i=0}^{k}\Big(1 - \frac{\mu}{L}\Big)^{k-i} = \Big(1 - \frac{\mu}{L}\Big)^{k+1}\big(f(x^0) - f^*\big) + \frac{3\delta L}{\mu}.$$

Данное неравенство приводит к таким выводам. С одной стороны мы видим, что величина, связанная с погрешностью δ, ограничена. Однако она может быть довольно немалой при большом значении числа обусловленности $\dfrac{L}{\mu}$. Это показывает также, что замена слагаемого $\Delta\|y - x\|$ в (4.11) на $\dfrac{\Delta^2 + \|y - x\|^2}{2}$ может привести к ухудшению оценки качества решения при достаточно большом $\dfrac{L}{\mu}$.

З а м е ч а н и е 5. Аналогично второй части доказательства теоремы 1 можно проверить, что трудоемкость итерации адаптивного алгоритма 2 сопоставима с трудоемкостью аналогичного неадаптивного метода.

З а м е ч а н и е 6. Предложенные в этом разделе подходы можно применять и для некоторых задач негладкой оптимизации (см. предыдущий раздел и определение 3), для которых целевая функция удовлетворяет (PL)-условию (в частности, μ-сильно выпукла). Если определяющий степень негладкости функции параметр Δ достаточно мал, то найденные оценки (4.6)–(4.10) (см. также замечание 4) позволяют сделать вывод о близкой к линейной скорости сходимости.

Заключение

В настоящей работе рассмотрены некоторые подходы к концепции неточной модели целевой функции в оптимизации, которые учитывают как погрешность задания целевого функционала, так и погрешность задания градиента. Предложены методы с адаптивным выбором

шага, а также адаптивной настройкой величины в оценке скорости сходимости, которая определяется упомянутыми погрешностями.

Сравнивая алгоритмы 1 и 2 между собой, отметим следующее. Преимущества алгоритма 1 (и его модификации из третьего раздела статьи) состоят в максимальной общности (метод можно использовать для широкого класса задач выпуклой оптимизации [2; 3], в том числе с условиями относительной гладкости [4]). Также, в отличие от алгоритма 2, для работы алгоритма 1 и использования найденной оценки скорости сходимости нет необходимости знать Δ (оценку неточности задания градиента). Как преимущества алгоритма 2 для функций, удовлетворяющих (PL)-условию, можно упомянуть близкую к линейной скорость сходимости и возможность использования метода на неограниченном допустимом множестве. Однако для оценок (4.6)–(4.10) необходимо знать верхнюю оценку величины Δ. Также существенно использована безусловность поставленной задачи. Оценка для алгоритма 1 в свою очередь проигрывает полученной для алгоритма 2 возможностью сколь угодно большого влияния погрешности градиента при $\gamma > 0$ для неограниченной области Q. Хорошо известно, что (PL)-условие заведомо верно в случае μ-сильной выпуклости целевой функции f относительно евклидовой нормы. Однако довольно хорошо известны примеры, когда нельзя быть уверенным даже в выпуклости $f(x)$, но (PL)-условие имеет место (см., например, разд. 4.3 из диссертации [8]). Это означает, что алгоритм 2 применим и для некоторых задач невыпуклой оптимизации. Интересно, что все рассмотренные методы применимы к некоторому классу задач негладкой оптимизации (см. определение 3).

В качестве актуальной задачи на будущее можно было бы выделить проблему построения так называемых ускоренных методов для рассмотренных классов задач. В частности, к ускоренным методам относят самые разные вариации так называемого быстрого градиентного метода (БГМ) (см., например, [1; 5; 8]). Для задач выпуклой гладкой оптимизации без погрешностей БГМ гарантирует лучшую оценку скорости сходимости по сравнению с (1.2). Известно также, что в сильно выпуклом случае использование ускоренных методов позволяет уменьшить знаменатель геометрической прогрессии, которая описывает скорость сходимости. Более того, неадаптивные ускоренные методы для релаксаций условия сильной выпуклости исследовались в [1]. Однако стоит отметить, что при наличии погрешностей ситуация становится уже менее тривиальной: в отличие от обычного градиентного метода возможно их накопление в итоговой оценке [6], либо же необходимо использовать довольно ограничительные условия на величины таких погрешностей [17]. Также пока не удалось предложить ускоренный метод, который применим в общем случае для относительно гладких задач [4]. Представляется интересной задача исследования применимости результатов настоящей работы для приближенного решения бесконечномерных задач, в частности, для некоторых типов линейных и нелинейных операторных уравнений.

Автор благодарит Александра Владимировича Гасникова, а также рецензента за полезные обсуждения и замечания.

СПИСОК ЛИТЕРАТУРЫ

1. **Necoara I., Nesterov Y. Glineur F.** Linear convergence of first order methods for non-strongly convex optimization // Math. Program. 2019. Vol. 175. P. 69–107. doi: 10.1007/s10107-018-1232-1 .
2. **Тюрин А. И., Гасников А. В.** Быстрый градиентный спуск для задач выпуклой минимизации с оракулом, выдающим (δ, L)-модель функции в запрошенной точке. // Журн. вычисл. математики и мат. физики. 2019. Т. 59, № 7. С. 1137–1150. doi: 10.1134/S0044466919070081 .
3. **Stonyakin F. S., Dvinskikh D., Dvurechensky P., Kroshnin A., Kuznetsova O., Agafonov A., Gasnikov A., Tyurin A., Uribe C. A., Pasechnyuk D., Artamonov S.** Gradient methods for problems with inexact model of the objective // Intern. Conf. on Mathematical optimization theory and operations research (MOTOR 2019): extended conference abstracts / eds. M. Khachay, Y. Kochetov, P. Pardalos. Cham: Springer, 2019. P. 97–114. (Lecture Notes in Computer Science; vol. 11548). doi: 10.1007/978-3-030-22629-9_8 .

4. **Lu H., Freund R. M., Nesterov Y.** Relatively smooth convex optimization by Firstorder methods, and applications. // SIAM J. Optim. 2018. Vol. 28, no 1. P. 333–354. doi: 10.1137/16M1099546 .
5. **Devolder O., Glineur F., Nesterov Yu.** First-order methods of smooth convex optimization with inexact oracle. // Math. Program. 2014. Vol. 146, no. 1–2. P. 37–75. doi: 10.1007/s10107-013-0677-5 .
6. **Devolder O.** Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization: PhD thesis. 2013. 320 p.
7. **Nesterov Yu.** Gradient methods for minimizing composite functions // Math. Program. 2013. Vol. 140, no. 1. P. 125–161. doi: 10.1007/s10107-012-0629-5 .
8. **Нестеров Ю. Е.** Алгоритмическая выпуклая оптимизация: дисс. . . . д-р физ.-мат. наук: 01.01.07. М.: МФТИ, 2013. 367 с.
9. **Karimi H., Nutini J., Schmidt M.** Linear convergence of πradient and ɜroximal-gradient methods under the Polyak–Lojasiewicz condition // Machine Learning and Knowledge Discovery in Databases / eds. B. Berendt etc. Cham: Springer, 2016. P. 795–811. (Lecture Notes in Computer Science; vol. 9851). doi: 10.1007/978-3-319-46128-1_50 .
10. **Mordukhovich B.** Variational analysis and generalized differentiation I, theory and examples. Berlin; Heidelberg: Springer-Verlag, 2006. 579 p. (Part of the Grundlehren der mathematischen Wissenschaften book series; vol 330). doi: 10.1007/3-540-31247-1 .
11. **Поляк Б. Т.** Градиентные методы минимизации функционалов // Журн. вычисл. математики и мат. физики. 1963. Том 3, № 4. С. 643–653. doi: 10.1016/0041-5553(63)90382-3 .
12. **Поляк Б. Т.** Введение в оптимизацию. М.: Наука, 1983. 384 с.
13. **Стонякин Ф. С.** Аналог квадратичной интерполяции для специального класса негладких функционалов и одно его приложение к адаптивному методу зеркального спуска // Динамические системы. 2019. Т. 9 (37), № 1. С. 3–16.
14. **Mordukhovich B. S., Nam N. M.** Applications of variational analysis to a generalized Fermat–Torricelli problem // J. Optim. Theory Appl. 2011. Vol. 148, no. 3. P. 431–454. doi: 10.1007/s10957-010-9761-7 .
15. **Немировский А. С., Юдин Д. Б.** Сложность задач и эффективность методов оптимизации. М.: Наука. Гл. редакция физ.-мат. литературы, 1979. 384 с.
16. **Nesterov Yu.** Universal gradient methods for convex optimization problems // Math. Program. Ser. A. 2015. Vol. 152, iss. 1-2. P. 381–404. doi: 10.1007/s10107-014-0790-0 .
17. **D'Aspremont A.** Smooth optimization with approximate gradient. // SIAM J. Optimi. 2008. Vol. 19, no. 3. P. 1171–1183. doi: 10.1137/060676386 .

Поступила 8.09.2019
После доработки 21.10.2019
Принята к публикации 28.10.2019

Стонякин Федор Сергеевич
канд. физ.-мат. наук, доцент
Крымский федеральный университет им. В. И. Вернадского
г. Симферополь
e-mail: fedyor@mail.ru

REFERENCES

1. Necoara I., Nesterov Y. Glineur F. Linear convergence of first order methods for non-strongly convex optimization. *Math. Program.*, 2019, vol. 175, pp. 69–107. doi: 10.1007/s10107-018-1232-1 .
2. Tyurin A. I., Gasnikov A. V. Fast gradient descent for convex minimization problems with an oracle issuing (δ, L)-model of the function at the requested point. *Comput. Math. Math. Phys.*, 2019, vol. 59, no. 7, pp. 1085–1097. doi: 10.1134/S0044466919070081 .
3. Stonyakin F. S., Dvinskikh D., Dvurechensky P., Kroshnin A., Kuznetsova O., Agafonov A., Gasnikov A., Tyurin A., Uribe C. A., Pasechnyuk D., Artamonov S. Gradient methods for problems with inexact model of the objective. In: M. Khachay, Y. Kochetov, P. Pardalos (eds.) Mathematical Optimization Theory and Operations Research (MOTOR 2019), *Lecture Notes in Computer Science*, vol. 11548, Cham: Springer, 2019, pp. 97–114. doi: 10.1007/978-3-030-22629-9_8 .
4. Lu H., Freund R. M., Nesterov Y. Relatively smooth convex optimization by Firstorder methods, and applications. *SIAM J. Optim.*, 2018, vol. 28, no 1, pp. 333–354. doi: 10.1137/16M1099546 .

5. Devolder O., Glineur F., Nesterov Yu. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 2014, vol. 146, no. 1-2, pp. 37–75. doi: 10.1007/s10107-013-0677-5 .
6. Devolder O. Exactness, Inexactness and Stochasticity in First-Order Methods for Large-Scale Convex Optimization, *PhD thesis*, 2013, 320 p.
7. Nesterov Yu. Gradient methods for minimizing composite functions. *Math. Program.*, 2013, vol. 140, no. 1, pp. 125–161. doi: 10.1007/s10107-012-0629-5 .
8. Nesterov Yu. E. *Algoritmicheskaya vypuklaya optimizatsiya* [Algorithmic convex optimization]. Doctor Sci. (Phys.-Math.) Dissertation. Moscow: Mosk. Phys.-Tech. Inst. (State University), 2013, 367 p.
9. Karimi H., Nutini J., Schmidt M. Linear convergence of gradient and proximal-gradient methods under the Polyak–Lojasiewicz condition. In: B. Berendt etc. (eds.), Machine Learning and Knowledge Discovery in Databases, *Lecture Notes in Computer Science,* vol. 9851, Cham: Springer, 2016, pp. 795–811. doi: 10.1007/978-3-319-46128-1_50 .
10. Mordukhovich B. *Variational analysis and generalized differentiation* I*, Theory and examples*, Part of the Grundlehren der mathematischen Wissenschaften book series, vol 330, Berlin, Heidelberg: Springer-Verlag, 2006, 579 p. doi: 10.1007/3-540-31247-1 .
11. Polyak B. T. Gradient methods for minimizing functionals. *Comput. Math. Math. Phys.*, 1963, vol. 3, no. 4, pp. 643–653. doi: 10.1016/0041-5553(63)90382-3 .
12. Polyak B. T. *Vvedenie v optimizaciju* [Introduction to Optimization]. Moscow: Nauka Publ., 1983, 384 p.
13. Stonyakin F. S. An analog of quadratic interpolation for a special class of non-smooth functionals and one of its applications to the adaptive method of mirror descent. *Dinamicheskie sistemy*, 2019, vol. 9 (37), no. 1, pp. 3–16 (in Russian).
14. Mordukhovich B. S., Nam N. M. Applications of variational analysis to a generalized Fermat–Torricelli problem. *J. Optim. Theory Appl.*, 2011, vol. 148, no. 3, pp. 431–454. doi: 10.1007/s10957-010-9761-7 .
15. Nemirovsky A. S., Yudin D. B. *Slozhnost' zadach i effektivnost' metodov optimizatsii* [The complexity of tasks and the effectiveness of optimization methods]. Moscow: Nauka Publ., 1979, 384 p.
16. Nesterov Yu. Universal gradient methods for convex optimization problems. *Math. Program. Ser. A*, 2015, vol. 152, iss. 1-2, pp. 381–404. doi: 10.1007/s10107-014-0790-0 .
17. D'Aspremont A. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 2008, vol. 19, no 3, pp. 1171–1183. doi: 10.1137/060676386 .

Received September 8, 2019
Revised October 21, 2019
Accepted October 28, 2019

Funding Agency: This work was supported by the Russian Science Foundation (project no. 18-71-00048).

Fedor Sergeevich Stonyakin, Cand. Sci. (Phys.-Math.), V.I. Vernadsky Crimean Federal University, Simferopol, Republic of Crimea, 295007 Russia, e-mail: fedyor@mail.ru .

Cite this article as: F. S. Stonyakin, Adaptation to inexactness for some gradient-type optimization methods, *Trudy Instituta Matematiki i Mekhaniki URO RAN*, 2019, vol. 25, no. 4, pp. 210–225 .