УДК 519.85

О некоторых работах Бориса Теодоровича Поляка по сходимости градиентных методов и их развитии*

© 2023 г. С. С. Аблаев[1,4], А. Н. Безносиков[1,2], А. В. Гасников[1,2,3], Д. М. Двинских[1,2,3], А. В. Лобанов[1,3], С. М. Пучинин[1], Ф. С. Стонякин[1,4]

[1]141700 Долгопрудный, М.о., Институтский пер., 9, НИУ МФТИ;
[2]127051, г. Москва, Б. Каретный переулок, д.19 стр. 1, ИППИ РАН;
[3]109004 Москва, ул. А. Солженицына, 25, ИСП РАН;
[4]295007 Республика Крым, г. Симферополь, просп. академика Вернадского 4, Крымский федеральный университет имени В.И. Вернадского;
e-mail: gasnikov@yandex.ru

Поступила в редакцию: 15.09.2023 г.
Переработанный вариант 16.12.2023 г.
Принята к публикации 17.11.2023 г.

Аннотация: В статье представлен обзор современного состояния субградиентных и ускоренных методов выпуклой оптимизации, в том числе при наличии помех и доступа к различной информации о целевой функции (значение функции, градиент, стохастический градиент, старшие производные). Для невыпуклых задач рассматривается условие Поляка–Лоясиевича и приводится обзор основных результатов. Рассматривается поведение численных методов при наличии острого минимума. Цель данного обзора – показать влияние работ Б.Т. Поляка (1935 – 2023) по градиентным методам оптимизации и их окрестностям на современное развитие численных методов оптимизации.

Ключевые слова: градиентный спуск, условие градиентного доминирования (Поляка–Лоясиевича), острый минимум, субградиентный метод Поляка–Шора, условие ранней остановки, метод тяжелого шарика Поляка, стохастический градиентный спуск

1. Введение

Данный обзор посвящен частичному разбору нескольких работ Б.Т. Поляка, о сходимости методов градиентного типа, которые на многие десятилетия определили развитие численных методов оптимизации. В частности, продолжают активно цитироваться и развиваться в настоящее время. Прежде всего, речь пойдет об этих работах [1—15].

Подчеркнем, что в данной статье не планируется описывать научный путь Бориса Теодоровича. Мы коснемся лишь пары десятков статей из более чем 250. Более подробно с научным путем Б.Т. Поляка можно познакомиться, например, по статьям [16; 17].

Структура обзора следующая. В разделе 2 описываются методы негладкой оптимизации и специальный способ адаптивного выбора шага (в литературе часто называется «шаг Поляка»). В частности, приводится субградиентный метод Поляка–Шора и вариант этого метода с переключением (для задач с функциональными ограничениями). Далее обсуждается вопрос о возможной линейной скорости сходимости таких методов, если минимум острый.

*Работа выполнена при финансовой поддержке Минобрнауки РФ (проект FSMG-2024-0011).

В разделе 3 излагаются методы гладкой оптимизации. Начинается изложение с описания условия градиентного доминирования, которое обобщает условие сильной выпуклости целевого функционала, не предполагая при этом даже выпуклости. Показывается, что при данном условии градиентный спуск глобально линейно сходится. Рассматриваются различные обобщения данного результата. В частности, на случай, когда градиент доступен лишь с заданным уровнем относительной неточности. Затем идет изложение метода условного градиента и некоторых современных результатов вокруг этого метода. В заключение данного раздела описывается метод тяжелого шарика Поляка, который породил линейку современных ускоренных методов.

В заключительном разделе 4 приводятся результаты Поляка–Цыпкина и Поляка–Юдицкого–Рупперта, в которых возникают различные формы центральной предельной теоремы для выхода алгоритма типа стохастического градиентного спуска. Далее приводятся неасимптотические результаты, в том числе и для ускоренных вариантов стохастического градиентного спуска. В частности, рассматриваются, так называемые, мультипликативные помехи, которые в современных исследованиях чаще называют условием сильного роста, не ассоциируя это с пионерскими работами Б.Т. Поляка с соавторами. В заключение рассматриваются рандомизированные безградиентные (также называемые, поисковые) методы. В условиях повышенной гладкости целевой функции обсуждается конструкция Поляка–Цыбакова, позволяющая строить хорошую модель производной по направлению целевой функции, исходя всего из двух проб.

2. Негладкая выпуклая оптимизация

История развития методов негладкой оптимизации начинается в 60-е годы прошлого века и достаточно подробно описана в работе [18]. Наряду с работами Н.З. Шора [19] важный вклад в развитие этой области принадлежит Б.Т. Поляку [11; 15].

2.1. Субградиентный метод Поляка–Шора

Если не оговорено иное, то далее в тексте рассматриваются задачи вида

$$f(x) \to \min_{x \in Q}, \tag{1}$$

где $f(x)$ – необязательно дифференцируемая выпуклая функция, Q – выпуклое замкнутое подмножество $\mathbb{R}^n$, $f^* = f(x_*) = \min_{x \in Q} f(x)$.

В данном разделе мы поговорим про субградиентные методы для негладкой оптимизации и вклад Бориса Теодоровича Поляка. Хорошо известно, что при минимизации недифференцируемых функций возникает ряд проблем: неприменим метод покоординатного спуска, а субградиент целевой функции не задаёт направление наискорейшего возрастания. В связи с этим Н. З. Шором был предложен субградиентный метод [20], являющийся прямым аналогом градиентного метода. Особенность идеи такого метода в том, что вместо градиента целевой функции в методе используется произвольный субградиент негладкой выпуклой функции. Рассмотрим случай, когда Q — замкнутое выпуклое подмножество $\mathbb{R}^n$ с евклидовой нормой $\|\cdot\|_2$. Пусть $B_2^n(x_*, R) = \{x \in \mathbb{R}^n : \ \|x - x_*\|_2 \leqslant R\}$. Будем всюду далее обозначать субградиент f (некоторый элемент субдифференциала $\partial f(x)$) в точке x как $\nabla f(x)$. Если функция f дифференцируема в точке x, то $\nabla f(x)$ —

её градиент. Итерация субградиентного метода при $h_k > 0$ имеет следующий вид

$$x^{k+1} = Pr_Q\{x^k - h_k \nabla f(x^k)\}, \quad \nabla f(x^k) \in \partial f(x^k), \tag{2}$$

где $\mathrm{Pr}_Q(y) := \arg\min_{x \in Q}\{\|y - x\|_2\}$ — оператор евклидова проектирования на множество Q. Одна из главных особенностей субградиентных методов состоит в том, что значения функции в этом методе могут не убывать монотонно с ростом количества итераций. Вообще, f не обязательно убывает вдоль направления $-\nabla f(x^k)$, обратного направлению субградиента в текущей точке. Однако оказывается, что при этом возможно гарантировать монотонное убывание расстояния от текущей точки до точки минимума. Вторая особенность — это выбор шага субградиентного метода. Если выбирать постоянный шаг, то метод может не сходиться. Действительно, пусть для функции одной переменной $f(x) = |x|$, $x_0 = -0.01$ и выбран постоянный шаг $h_k = 0.02$. Тогда итеративная последовательность метода (2) будет состоять всего из двух точек -0.01 и 0.01 и не будет сходимости к точке минимума 0.

Интересный подход к выбору шага в субградиентном методе предложен Б.Т. Поляком. Он предложил при выборе шага использовать степень близости значения функции в текущей точке к минимальному. Это вполне возможно, если известно искомое минимальное значение функции f^*. Например, если систему совместных линейных уравнений

$$\langle a_i, x\rangle = b_i, \quad i = 1, \ldots, n, \quad x \in \mathbb{R}^n,$$

свести к минимизации функции

$$f(x) = \sum_{i=1}^{n} \left|\langle a_i, x\rangle - b_i\right|,$$

то $f^* = 0$. Также f^* бывает известно в геометрических задачах: проекция точки на множество, нахождение общей точки системы множеств. Используя f^* можно построить адаптивный вариант шага (впервые он предложен в [11]), не содержащий таких параметров задачи, как константа Липшица целевой функции или расстояние от точки старта до множества решений:

$$h_k = \frac{f(x^k) - f^*}{\|\nabla f(x^k)\|_2^2}. \tag{3}$$

Такой шаг принято называть шагом Б.Т. Поляка. Для субградиентного метода с таким шагом известен следующий результат о сходимости [2].

Теорема 1. *Пусть $f(x)$ — выпуклая на $\mathbb{R}^n$ функция, множество точек минимума X_* которой не пусто. Тогда в методе (2) с шагом (3) $x^k \to x_* \in X_*$. При этом $\lim\limits_{k\to\infty} \sqrt{k}(f(x^k) - f^*) = 0$.*

Доказательство. Как известно, для всякой точки минимума $x_* \in X_*$ верны неравенства

$$2h_k(f(x^k) - f(x_*)) \le 2h_k\langle \nabla f(x^k), x^k - x_*\rangle \le$$

$$\le h_k^2\|\nabla f(x^k)\|_2^2 + \|x^k - x_*\|_2^2 - \|x^{k+1} - x_*\|_2^2.$$

Поэтому

$$\|x^{k+1} - x_*\|_2^2 \le h_k^2\|\nabla f(x^k)\|_2^2 - 2\,h_k(f(x^k) - f(x_*)) + \|x^k - x_*\|_2^2 = \tag{4}$$

$$= \frac{(f(x^k) - f(x_*))^2}{\|\nabla f(x^k)\|_2^2} - \frac{2\,(f(x^k) - f(x_*))^2}{\|\nabla f(x^k)\|_2^2} + \|x^k - x_*\|_2^2 =$$

$$= -\frac{(f(x^k) - f(x_*))^2}{\|\nabla f(x^k)\|_2^2} + \|x^k - x_*\|_2^2. \tag{5}$$

Таким образом,

$$\frac{f(x^k) - f^*}{\|\nabla f(x^k)\|_2} \to 0.$$

Более того, неравенство $\|x^k - x_*\|_2 \leq \|x^0 - x_*\|_2$ означает, что последовательность x^k ограничена, и тогда $\|\nabla f(x^k)\|_2 \leq M$. Поэтому $f(x^k) \to f^*$. Следовательно, найдётся подпоследовательность $x^{k_l} \to x_*$. Итак, получаем, что $\|x^k - x_*\|_2$ монотонно убывает, а $\|x^{k_l} - x_*\|_2 \to 0$. Отсюда $x^k \to x_*$. Ввиду (5) получаем

$$\sum_{k=0}^{\infty} \frac{(f(x^k) - f^*)^2}{\|\nabla f(x^k)\|_2^2} < \infty,$$

а из ограниченности $\|\nabla f(x^k)\|_2$ следует $\sum\limits_{k=0}^{\infty} (f(x^k) - f^*)^2 < \infty$. Если предположить, что $\lim\limits_{k\to\infty} \sqrt{k}(f(x^k) - f^*) > 0$, то $f(x^k) - f^* > \frac{a}{\sqrt{k}}$ для достаточно больших k, что противоречит условию $\sum\limits_{k=0}^{\infty} (f(x^k) - f^*)^2 < \infty$. Итак, $\lim\limits_{k\to\infty} \sqrt{k}(f(x^k) - f^*) = 0$. □

Предыдущий результат означает, что для достижения точности $\varepsilon > 0$ решения задачи по функции гарантированно достаточно $O\left(\frac{1}{\varepsilon^2}\right)$ итераций. Данная оценка скорости сходимости неулучшаема на классе минимизационных задач с выпуклыми липшицевыми (как гладкими, так и негладкими) целевыми функциями. Хотя известны и другие подходы к выбору шага для субградиентного метода. Например, если при $Q = \mathbb{R}^n$ предположить (здесь и всюду далее $R \geqslant \left\|x^0 - x_*\right\|_2$), что

$$\|\nabla f(x)\|_2 \leqslant M \text{ для всякого } x \in B_2^n(x_*, R\sqrt{2}), \tag{6}$$

и выбрать шаг субградиентного метода, а также точку выхода следующим образом:

$$h = \frac{R}{M\sqrt{N}}, \quad \overline{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k, \tag{7}$$

то будет выполняться неравенство

$$f\left(\overline{x}^N\right) - f(x_*) \leqslant \frac{MR}{\sqrt{N}}. \tag{8}$$

Неравенство (8) означает, что при выборе

$$N = \frac{M^2R^2}{\varepsilon^2}, \quad h = \frac{\varepsilon}{M^2} \tag{9}$$

будет достигаться оценка $f\left(\overline{x}^N\right) - f(x_*) \leqslant \varepsilon$, которая оптимальна с точностью до умножения на константу. Действительно, известно, что точная нижняя оценка на классе задач выпуклой оптимизации с условием (6) для методов первого порядка вида

$$x^{k+1} \in x^0 + \operatorname{span}\left\{\nabla f(x^0), ..., \nabla f(x^k)\right\}, \tag{10}$$

где для всех $j = 0, 1, ..., k$ верно $\nabla f(x^j) \in \partial f(x^j)$, имеет вид [21]:

$$f\left(x^N\right) - f(x_*) \geqslant \frac{MR}{\sqrt{N+1}}. \tag{11}$$

Представляется, что интерес выбора шага Б.Т. Поляка для субградиентного метода в том, что он в каждой точке позволяет учесть динамику значений целевой функции и не содержит параметров типа константы Липшица целевой функции или оценки расстояния от начальной точки до множества точных решений задачи. Такого типа процедуры выбора шагов в оптимизационных методах часто называют адаптивными. Разным подходам к адаптивным процедурам при выборе шагов оптимизационных методов посвящаются всё новые работы [22—24]. В этом плане можно отметить многочисленные исследования в области универсальных градиентных методов [25; 26], адаптивных методов типа AdaGrad, да и разные модификации шага Б.Т. Поляка в детерминированном и стохастическом случае [22; 27]. Важно, что, помимо процедуры выбора шага, Б.Т. Поляк ещё выделил и класс задач с острым минимумом [11], для которого этот шаг позволил доказать результат о сходимости субградиентного метода со скоростью геометрической прогрессии. Далее поговорим об этом результате и его развитии в современных работах.

2.2. Острый минимум и линейная скорость сходимости субградиентного метода с шагом Б.Т. Поляка

Как видим, в описанных выше результатах оценки скорости сходимости субградиентного метода сублинейны. Получить линейную сходимость субградиентного метода можно лишь с помощью дополнительных предположений. Например, линейная скорость сходимости возможна для методов типа секущей гиперплоскости, применимых к задачам малой или умеренной размерности. Что касается задач большой размерности, то сходимость со скоростью геометрической прогрессии может позволить получить дополнительное предположение об остром минимуме

$$f(x) - f^* \geqslant \alpha \min_{x_* \in X_*} \|x - x_*\|_2, \tag{12}$$

где X_* — множество точек минимума функции f на множестве Q, $\alpha > 0$ — некоторое фиксированное положительное число. В частности, условие (12) верно для задачи евклидова проектирования точки x на выпуклый компакт $X_* \subset Q$, причем $f^* = 0$. Условие острого минимума было введено Б.Т. Поляком в [11]. Рассмотрим субградиентный метод (2) с шагом Б. Т. Поляка [11] (3) для задачи минимизации f.

Теорема 2. *Пусть f — выпуклая функция и для задачи минимизации f с острым минимумом используется алгоритм (2) с шагом Б. Т. Поляка (3). Тогда после k итераций алгоритма (2) верно неравенство*

$$\min_{x_* \in X_*} \|x^{k+1} - x_*\|_2^2 \leq \prod_{i=0}^{k} \left(1 - \frac{\alpha^2}{\|\nabla f(x^i)\|_2^2}\right) \min_{x_* \in X_*} \|x^0 - x_*\|_2^2.$$

Доказательство. Согласно (5) и условию острого минимума имеем:

$$\min_{x_* \in X_*} \|x^{k+1} - x_*\|_2^2 \leq -\frac{\alpha^2}{\|\nabla f(x^k)\|_2^2} \min_{x_* \in X_*} \|x^k - x_*\|_2^2 + \min_{x_* \in X_*} \|x^k - x_*\|_2^2 = \tag{13}$$

$$= \Big(1 - \frac{\alpha^2}{\|\nabla f(x^k)\|_2^2}\Big) \min_{x_* \in X_*} \|x^k - x_*\|_2^2.$$

Далее, получаем цепочку неравенств

$$\min_{x_* \in X_*} \|x^{k+1} - x_*\|_2^2 \le \Big(1 - \frac{\alpha^2}{\|\nabla f(x^k)\|_2^2}\Big) \min_{x_* \in X_*} \|x^k - x_*\|_2^2 \le \tag{14}$$

$$\le \Big(1 - \frac{\alpha^2}{\|\nabla f(x^k)\|_2^2}\Big)\Big(1 - \frac{\alpha^2}{\|\nabla f(x^{k-1})\|_2^2}\Big) \min_{x_* \in X_*} \|x^{k-1} - x_*\|_2^2 \le \ldots \le$$

$$\le \prod_{i=0}^{k} \Big(1 - \frac{\alpha^2}{\|\nabla f(x^i)\|_2^2}\Big) \min_{x_* \in X_*} \|x^0 - x_*\|_2^2. \tag{15}$$

□

Следствие 1. *Если в условиях предыдущей теоремы допустить, что f удовлетворяет условию Липшица с константой $M > 0$ (т.е. все нормы субградиентов f равномерно сверху ограничены этой константой), то можно утверждать сходимость алгоритма (2) со скоростью геометрической прогрессии*

$$\min_{x_* \in X_*} \|x^{k+1} - x_*\|_2^2 \le \Big(1 - \frac{\alpha^2}{M^2}\Big)^{k+1} \min_{x_* \in X_*} \|x^0 - x_*\|_2^2.$$

2.3. Субградиентные схемы с переключениями для задач с ограничениями

Б. Т. Поляк внёс существенный вклад и в изучение задач математического программирования. Ему принадлежит идея так называемых схем с переключениями по продуктивным и непродуктивным шагам [28]. Общая идея подхода заключается в следующем: если в текущей точке значение ограничения достаточно хорошее, то спуск выполняем по целевой функции, а в противном случае — по функции ограничения. Такого типа подходам, которые интересны ввиду малых затрат памяти на итерациях, посвящаются всё новые работы как для выпуклых задач большой и сверхбольшой размерности [29—32], так и для некоторых классов невыпуклых задач [29]. В последние годы некоторыми из авторов статьи были исследованы адаптивные субградиентные методы с переключениями для липшицевых задач выпуклого программирования, в том числе и для некоторых нелипшицевых и/или квазивыпуклых целевых функций [30; 33—35]. Достаточно полное исследование стохастического варианта субградиентного метода с переключениями по продуктивным и непродуктивным шагам имеется в работе [36].

В этом пункте приводится результат о сходимости субградиентных методов со скоростью геометрической прогрессии для субградиентных схем с переключениями по продуктивным и непродуктивным шагам в случае выпуклых задач с ограничениями в виде неравенств. Случай квазивыпуклых задач с квазивыпуклыми ограничениями исследован в [35]. Будем рассматривать задачу с функциональными ограничениями вида

$$\begin{cases} \min f(x), \\ x \in Q,\ g(x) \le 0, \end{cases} \tag{16}$$

где $f(x)$ и $g(x)$ — липшицевы функции. Всюду далее будем считать, что задача (16) разрешима.

Algorithm 1 Адаптивный субградиентный метод для выпуклой целевой функции.

Require: $\delta > 0, M_g > 0, x^0, \theta_0 : \theta_0^2 \geq \frac{1}{2}\|x_* - x^0\|_2^2$, множество Q.

$I =: \emptyset$
$N \leftarrow 0$
repeat
  if $g\left(x^N\right) \leq \delta M_g$ **then**
    $h_N^f = \frac{\delta}{\|\nabla f(x^N)\|_2^2}$,
    $x^{N+1} = \mathrm{Pr}_Q\left(x^N - h_N^f \nabla f(x^N)\right)$, // *«продуктивные шаги»*
    $N \to I$,
  else
    $h_N^g = \frac{\delta}{\|\nabla g(x^N)\|_2}$,
    $x^{N+1} = \mathrm{Pr}_Q\left(x^N - h_N^g \nabla g(x^N)\right)$, // *«непродуктивные шаги»*
  end if
  $N \leftarrow N + 1$,
until $\frac{2\theta_0^2}{\delta^2} \leq \sum\limits_{k \in I} \frac{1}{\|\nabla f(x^k)\|_2^2} + N - |I|$.

Ensure: $\widehat{x} := \arg\min_{x^k,\, k \in I} f(x^k)$.

Рассмотрим теперь алгоритм 1, где I и J — множества индексов продуктивных и непродуктивных шагов соответственно, а $|I|$ и $|J|$ — мощности этих множеств.

Покажем пару примеров, несколько поясняющих смысл использования таких схем для задач с ограничениями.

Пример 1. *Рассмотрим случай задачи с несколькими ограничениями*

$$g(x) = \max\{g_1(x), ..., g_m(x)\}. \tag{17}$$

Можно сэкономить время работы алгоритма за счет рассмотрения не всех функциональных ограничений на непродуктивных шагах. То есть на непродуктивных шагах вместо субградиента ограничения $g(x)$ можно рассматривать субградиент любого из функционалов $g_m(x)$, для которого верно $g_m(x^k) > \varepsilon$. Другие ограничения при этом можно игнорировать. Легко проверить, что результат о сходимости алгоритма 1 при этом сохранится. Аналогичное замечание можно сделать и про иные подходы указанного типа [33].

Пример 2. *Интересным примером приложений может быть использование схем с переключениями к задачам проектирования механических конструкций, сводящихся к минимизации функций max-типа с разреженной матрицей A [32; 37]*

$$\max_y \langle f, y\rangle : \quad g(y) := \max_{1 \leqslant i \leqslant d} (\pm\langle a_i, y\rangle - 1) = \max_{1 \leq i \leq 2d} g_i(y) \leqslant 0. \tag{18}$$

Основное преимущество применения субградиентного метода с переключениями для задач выпуклого программирования заключается в том, что сложность одной итерации сильно понижается за счёт разреженности векторов a_i, f. Из этого следует, что на каждом шаге

$$y^{k+1} = y^k - h_g \cdot \nabla g(y^k) \qquad \textit{или} \qquad y^{k+1} = y^k + h_f \cdot f$$

в векторе y обновляется не больше чем $r = O(1)$ элементов. Из того, что в каждом узле встречаются не более чем s стержней, следует, что обновление y влечёт за собой обновление максимум sr скалярных произведений $\langle a_i, y\rangle$.

Теорема 3. *После остановки алгоритма 1 для всякого M_g-липшицева квазивыпуклого ограничения g верно $f(\widehat{x}) - f(x_*) \leq \delta$ и $g(\widehat{x}) \leq \delta M_g$. При этом в случае M_f-липшицевой функции f достаточное количество итераций для выполнения критерия остановки алгоритма 1 оценивается следующим образом:*

$$N \geq \frac{2\theta_0^2 \max\{1, M_f^2\}}{\delta^2}.$$

Замечание 1. *Заметим, что на практике для реализации алгоритма (1) знание константы Липшица M_f функции $f(x)$ необязательно, достаточно остановить алгоритм после выполнения критерия остановки.*

2.4. Аналог условия острого минимума для задач с ограничениями

Для выпуклых (в том числе негладких) липшицевых задач можно доказать сходимость субградиентного метода со скоростью геометрической прогрессии при дополнительном предположении о выполнении условия острого минимума [35]. В частности [2], острым минимумом будет обладать задача вида

$$\begin{cases} \min\langle c, x\rangle; & x \in \mathbb{R}^n, \\ Ax \leq b; & b \in \mathbb{R}^m, \end{cases} \tag{19}$$

где c — n-мерный вектор, A — матрица порядка $m \times n$.

Если для линейных задач (19) возможно обойтись использованием обычного условия острого минимума (12), то в общем случае для нелинейных задач это уже не так.

Для такой постановки (16) будем использовать следующую вариацию понятия острого минимума [35; 38].

Определение 1. *Будем говорить, что для задачи вида (16) выполняется условие острого минимума («условный» острый минимум), если при некотором $\alpha > 0$ для всех x справедливо неравенство*

$$\max\{f(x) - f^*, g(x)\} \geq \alpha \min_{x_* \in X_*} \|x - x_*\|_2. \tag{20}$$

Смысл такого подхода следующий. Поскольку задача (16) с ограничениями в виде неравенств, то в тех точках x, где эти неравенства нарушены, возможно $f(x) \leq f^*$, и тогда выполнение неравенства (20) возможно за счёт положительного g (очевидный пример — когда в качестве ограничения выбирается функция расстояния от точки до допустимого множества). В ситуации же $g(x) \leq 0$ потенциально возможна ситуация, когда неравенство $f(x) - f^* \geq \alpha \min\limits_{x_* \in X_*} \|x - x_*\|_2$ справедливо, а на всём допустимом множестве было бы неверным. Рассмотрим некоторые примеры таких задач с «условным» острым минимумом.

Пример 3. *Специальный вариант линейной задачи [38]*

$$\min -x_1, \tag{21}$$

$$\rho \cos\left(\frac{j\pi}{10}\right) x_1 + \rho \sin\left(\frac{j\pi}{10}\right) x_2 \le \rho, \quad j = 0, 1, \ldots, 19. \tag{22}$$

Как известно [38], в этой задаче точка минимума $x_* = (1, 0)$, а оптимальное значение $f^* = -1$. Отдельно целевая функция $-x_1$, вообще говоря, не удовлетворяет условию острого минимума (12), поскольку зависит только от одной переменной из двух. Но при наличии ограничений (22) для этой задачи выполнен «условный» вариант острого минимума (20) при $\alpha = \frac{\rho}{2}$ [38]. Выбор масштабирующего коэффициента ρ может влиять на значение параметра такого острого минимума.

Пример 4. *Рассмотрим задачу*

$$\begin{cases} \min\left\{f(x_1, x_2) = |x_1| + |x_2|\right\}, \\ g(x_1, x_2) = \max\{\varepsilon - x_1, \varepsilon - x_2\} \le 0, \end{cases}$$

где $\varepsilon > 0$. В этом примере целевая функция удовлетворяет условию острого минимума (12). Действительно, $f(x) = |x_1| + |x_2|$, $f^* = 0$, а $x_* = (0, 0)$, и неравенство $|x_1| + |x_2| \ge \alpha \|x\|_2$ выполняется, например, при $\alpha = 1$. Однако функциональные ограничения смещают точку минимума $x_* = (\varepsilon, \varepsilon)$, и условие острого минимума нарушается. А условному острому минимуму эта задача удовлетворяет, учитывая, что $x_* = (\varepsilon, \varepsilon)$, а минимальное значение функции $f^* = 2\varepsilon$.

Пример 5. *Задача классификации с ограничениями [38]. Рассмотрим множество, состоящее из n пар: $\mathcal{D} = \{(a_i, b_i)\}_{i=1}^n$, где $a \in \mathbb{R}^p$ — вектор признаков, а $b_i \in \{1, -1\}$ — множество меток при $i = 1, 2, \ldots, n$. Пусть $\mathcal{D}_M \in \mathbb{R}^p$, $\mathcal{D}_F \in \mathbb{R}^p$ — два типа признаков. Мы хотим найти линейный классификатор $x \in \mathbb{R}^p$, который не только минимизирует функцию потерь, но и правильным образом обрабатывает каждый элемент из множеств $\mathcal{D}_F$ и $\mathcal{D}_M$. Указанная задача сводится к следующей постановке:*

$$\min_x \frac{1}{n} \sum_{i=1}^{n} \max_x \{0, 1 - b_i a_i^T x\},$$

$$\frac{1}{n_F} \sum_{a \in \mathcal{D}_F} \sigma(a^T x) \ge \frac{\kappa}{n_M} \sum_{a \in \mathcal{D}_M} \sigma(a^T x),$$

$$\frac{1}{n_M} \sum_{a \in \mathcal{D}_M} \sigma(a^T x) \ge \frac{\kappa}{n_F} \sum_{a \in \mathcal{D}_F} \sigma(a^T x),$$

где $\kappa \in (0, 1]$ — константа, n_M и n_F — количество экземпляров из $\mathcal{D}_M$ и $\mathcal{D}_F$ соответственно и $\sigma(a^T x) := \max\{0, \min\{1, \{0.5 + a^T x\}\} \in [0, 1]$ — вероятность присваивания метки $+1$ предсказанию a.

Построим схему рестартов алгоритма 1 (алгоритм 2) в предположении, что верно условие (20).

Algorithm 2 Рестарты алгоритма 1.

Require: $\varepsilon > 0, \alpha > 0, M_g > 0, x^0, \theta_0 : \theta_0^2 \geq \frac{1}{2}\|x_* - x^0\|_2^2$, множество Q.

1: Set $p = 1$.
2: **repeat**
3: $\quad \widehat{x}^p$ — результат работы алгоритма 1 с параметрами δ_p, θ_p, x^0, где
4: $\quad x^0 = \widehat{x}^p$,
5: $\quad \theta_p = \frac{1}{\sqrt{2^p}}\theta_0$,
6: $\quad \delta_p = \frac{\alpha\theta_p}{\sqrt{2}\max\{1, M_g\}}$.
7: $\quad$ Set $p = p + 1$.
8: **until** $p > \left\lceil 2\log_2 \frac{\theta_0}{\varepsilon} \right\rceil$.

Ensure: x^p.

Для алгоритмов 1 и 2 верна следующая

Теорема 4. *[35] Пусть $f(x)$ и $g(x)$ — липшицевы функции с константами M_f и M_g соответственно, удовлетворяющие условию (20), и известна константа $\theta_0 > 0$ такая, что $2\theta_0^2 \geq \|x_* - x^0\|_2^2$. Тогда для алгоритма 1 можно подобрать параметр $\delta > 0$ так, чтобы после*

$$\left\lceil \frac{4}{\alpha^2} \max\left\{1, M_f^2\right\} \max\left\{1, M_g^2\right\} \right\rceil$$

итераций было выполнено неравенство $\min\limits_{x_* \in X_*} \|\widehat{x} - x_*\|_2 \leq \frac{1}{\sqrt{2}}\theta_0$. После $p - 1$ запусков алгоритма 2 (рестартов алгоритма 1) имеем

$$\min_{x_* \in X_*} \|\widehat{x}^{p-1} - x_*\|_2 \leq \frac{1}{\sqrt{2^p}}\theta_0.$$

Тогда для достижения ε-точного решения вида

$$\min_{x_* \in X_*} \|\widehat{x}^{p-1} - x_*\|_2 \leq \varepsilon$$

достаточное количество обращений к субградиенту f или g можно оценить как

$$\left\lceil \frac{4}{\alpha^2} \max\left\{1, M_f^2\right\} \max\left\{1, M_g^2\right\} \right\rceil \left\lceil 2\log_2 \frac{\theta_0}{\varepsilon} \right\rceil.$$

2.5. Субградиентные методы для слабо выпуклых и относительно слабо выпуклых задач с острым минимумом

Введённый Б.Т. Поляком класс выпуклых негладких оптимизационных задач с острым минимумом интересен тем, что на нём субградиентный метод имеет неплохие вычислительные гарантии (сходится со скоростью геометрической прогрессии). Однако представляется, что условие острого минимума достаточно существенно сужает рассматриваемый класс выпуклых задач. В этой связи интересно отметить наблюдение многих статей последних лет [39—42] о том, что многие возникающие в приложениях слабо выпуклые задачи имеют острый минимум. Это касается разных типов задач нелинейной регрессии, восстановления фазы, восстановления матрицы (matrix recovery) и др. Класс слабо выпуклых оптимизационных задач интересен, ему посвящаются всё новые работы

[29; 39; 43—47], но без дополнительных предположений скоростные гарантии достаточно пессимистичны. Однако в случае острого минимума недавно доказаны результаты о сходимости субградиентного метода со скоростью геометрической прогрессии при условии достаточной близости начальной точки к точному решению [39]. При этом заметим, что в одном из этих результатов используется способ выбора шага, предложенный Б.Т. Поляком. Немного расскажем об этом направлении.

Как и ранее, рассматриваем задачи минимизации вида (1). Напомним, что функция $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ (аналогично и для функций $f : Q \longrightarrow \mathbb{R}$) называется μ-слабо выпуклой ($\mu \geqslant 0$), если функция $x \longrightarrow f(x) + \frac{\mu}{2}\|x\|_2^2$ выпукла. Для недифференцируемых μ-слабо выпуклых функций f под субдифференциалом $\partial f(x)$ в точке x можно понимать (см. [39] и цитируемую там литературу) множество всех векторов $\nabla f(x) \in \mathbb{R}^n$, удовлетворяющих неравенству

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + o(\|y - x\|_2) \quad \text{при} \quad y \to x. \tag{23}$$

Известно [39], что все векторы-субградиенты $\nabla f(x) \in \mathbb{R}^n$ из (23) автоматически удовлетворяют неравенству

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle - \frac{\mu}{2}\|y - x\|_2^2 \quad \forall\, x, y \in \mathbb{R}^n,\ \nabla f(x) \in \partial f(x). \tag{24}$$

Можно проверить, что слабо выпуклые функции локально липшицевы, и поэтому в качестве субградиентов можно использовать произвольный вектор из субдифференциала Кларка (см., например, [43]).

Известно немало примеров прикладных слабо выпуклых задач, которые не являются выпуклыми. Так, слабо выпуклыми будут задачи вида (см., например, [39]):

$$\min_x f(x) := h(c(x)),$$

где $h : \mathbb{R}^m \longrightarrow \mathbb{R}$ выпукло и M-липшицево, а $c : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ является C^1-гладким отображением с β-липшицевой матрицей Якоби. Нетрудно проверить, что при указанных допущениях f — $M\beta$-слабо выпукла. В частности, в указанный класс задач входят задачи нелинейной регрессии с $h(x) = \|x\|_1$, где $\|\cdot\|_1$ — 1-норма.

Интересно, что слабая выпуклость позволяет существенно расширить класс задач с условием острого минимума, предложенного Б.Т. Поляком в [11]. Приведём ещё пару достаточно популярных в современных приложениях примеров слабо выпуклых задач с острым минимумом.

Пример 6 (Задача восстановления фазы)**.** *Восстановление фазы — распространенная вычислительная задача, имеющая приложения в различных областях, таких как визуализация, рентгеновская кристаллография и обработка речи [40; 41; 45].*

Она сводится к задаче минимизации следующей функции:

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} \left| \langle a_i, x \rangle^2 - b_i \right|, \tag{25}$$

где $a_i \in \mathbb{R}^n$ и $b_i \in \mathbb{R}$ заданы для каждого $i = 1, \ldots, m$. Данная целевая функция имеет вид $f(x) := h(c(x)) \to \min_x$, где $h\colon \mathbb{R}^m \to \mathbb{R}$ — выпуклая и M-липшицева функция, а $c\colon \mathbb{R}^n \to \mathbb{R}^m$ — это C^1-гладкое отображение с β-липшицевым отображением Якоби. Как отмечено выше, такие функции слабо выпуклы. В данном случае функция (25) μ-слабо выпукла [44] при

$$\mu = 2 \max_{1 \leq i \leq m} \|a_i\|_2^2. \tag{26}$$

Более того, при соответствующей модели шума в измерениях функция имеет острый минимум ([40], Proposition 3).

Пример 7. *[42]*

$$\text{minimize}_{U\in\mathbb{R}^{n\times r}}\left\{\xi(U) := \frac{1}{m}\|y - A(UU^T)\|_2^2\right\}. \tag{27}$$

Рассмотрим робастную задачу восстановления матрицы небольшого ранга *[42], в которой измерения искажены* всплесками. *В частности, мы предполагаем, что*

$$y = A(X^*) + s^*, \tag{28}$$

где $s^* \in \mathbb{R}^m$ — вектор всплесков-возмущений, у которого небольшая часть элементов имеет произвольную величину, а остальные элементы равны нулю. Кроме того, множество ненулевых элементов предполагается неизвестным.

Хорошо известно, что ℓ_2-функция потерь чувствительна к возмущениям, что приводит к недостаточной эффективности постановки (27) для восстановления базовой матрицы низкого ранга. Глобальные минимумы ξ в (27) отклоняются от базовой матрицы низкого ранга из-за всплесков, и большая доля всплесков приводит к большему возмущению. Напротив, функция потерь ℓ_1 более устойчива к выбросам и широко используется для обнаружения выбросов [48—50]. Это привело к идее использовать функцию потерь ℓ_1 и факторизованное представление матричной переменной для решения надежной задачи восстановления малоранговой матрицы :

$$\text{minimize}_{U\in\mathbb{R}^{n\times r}}\left\{f(U) := \frac{1}{m}\|y - A(UU^T)\|_1\right\}. \tag{29}$$

Известно [42], что эта целевая функция слабо выпуклая и обладает острым минимумом.

Отметим, что для слабо выпуклых задач характерны многие проблемы сходимости вычислительных процедур, свойственные общей невыпуклой ситуации. Например, нет возможности гарантировать сходимость градиентного метода даже к локальному минимуму.

Пример 8. *Действительно [51], пусть*

$$f(x) \equiv f(x^{(1)}, x^{(2)}) = \frac{1}{2}\left(x^{(1)}\right)^2 + \frac{1}{4}\left(x^{(2)}\right)^4 - \frac{1}{2}\left(x^{(2)}\right)^2.$$

Это слабо выпуклая задача, поскольку добавление к f половины квадрата нормы x приводит к выпуклой функции. Градиент целевой функции имеет вид

$$\nabla f(x) = \left(x^{(1)}, \left(x^{(2)}\right)^3 - x^{(2)}\right)^T,$$

откуда следует, что существуют только три точки, которые могут претендовать на локальный минимум:
$x_1^* = (0,0),\ x_2^* = (0,-1),\ x_3^* = (0,1)$. *Вычисляя матрицу Гессе*

$$\nabla^2 f(x) = \begin{pmatrix} 1 & 0 \\ 0 & 3\left(x^{(2)}\right)^2 - 1 \end{pmatrix},$$

заключаем, что x_2^* *и* x_3^* *являются точками изолированного локального минимума, в то время как* x_1^* *есть только* стационарная точка *нашей функции. Действительно,* $f(x_1^*) = 0$ *и* $f(x_1^* + \varepsilon e_2) = \left(\frac{\varepsilon^4}{4}\right) - \left(\frac{\varepsilon^2}{2}\right) < 0$ *при достаточно малых* ε*. Рассмотрим траекторию градиентного метода, начинающуюся в точке* $x_0 = (1, 0)$*. Обратим внимание на то, что вторая координата этой точки* 0*, поэтому вторая координата для* $\nabla f(x_0)$ *также* 0*. Следовательно, вторая координата точки* x_1 *равна нулю и т. д. Таким образом, вся последовательность точек, образованная градиентным методом, будет иметь нулевую вторую координату, что означает сходимость этой последовательности к* x_1^**.*

Теорема 5. [39] *Пусть* f *μ-слабо выпукла и имеет α-острый минимум ($\alpha,\ \mu > 0$), а точка* x_0*:* $\min\limits_{x_* \in X_*} \|x^0 - x_*\|_2 \leq \frac{\alpha\gamma}{\mu}$ *для некоторого фиксированного* $\gamma \in (0; 1)$*. Тогда для метода* (2) *с шагом* (3) *верно неравенство*

$$\min_{x_* \in X_*} \|x^{k+1} - x_*\|_2^2 \leq \prod_{i=0}^{k} \left(1 - \frac{\alpha^2(1-\gamma)}{\|\nabla f(x^i)\|_2^2}\right) \min_{x_* \in X_*} \|x^0 - x_*\|_2^2. \tag{30}$$

Следствие 2. *Если в условиях предыдущей теоремы допустить, что* f *удовлетворяет условию Липшица с константой* $M > 0$*, то можно утверждать сходимость алгоритма* (2) *со скоростью геометрической прогрессии*

$$\min_{x_* \in X_*} \|x^{k+1} - x_*\|_2^2 \leq \left(1 - \frac{\alpha^2(1-\gamma)}{M^2}\right)^{k+1} \min_{x_* \in X_*} \|x^0 - x_*\|_2^2.$$

Отметим, что в условиях теоремы 5 нулевой (суб)градиент возможен только в точке $x^k \in X_*$, поскольку справедливо следующее утверждение.

Предложение 1 (Окрестность множества X_* не имеет стационарных точек)**.** *Задача минимизации* μ*-слабо выпуклой функции* f *с* α*-острым минимумом не имеет стационарных точек* x*, удовлетворяющих условию*

$$0 < \min_{x_* \in X_*} \|x - x_*\|_2 < \frac{2\alpha}{\mu}. \tag{31}$$

3. Гладкая оптимизация

Пожалуй, основным атрибутом гладкой оптимизации является наличие моментного ускорения, роль которого, по-видимому, впервые была отмечена в 1964 году в работе Б.Т. Поляка [9]. Отметим, что в этой работе (как и в работе по негладким методам 1969 года [11]) были заложены основы разработки численных методов оптимизации на базе непрерывных аналогов (дифференциальных уравнений, порождающих при дискретизации итерационный процесс). Впоследствии аналогичным образом неоднократно разрабатывались новые численные методы оптимизации [52—55]. Однако наличие гладкости дает возможность не только ускорения скорости сходимости. В данном разделе, следуя работам Б.Т. Поляка, в основном 60-х годов прошлого века, продемонстрировано, какие дополнительные возможности приобретаются при наличии гладкости (липшицевости градиента целевой функции).

3.1. Условие градиентного доминирования (Поляка–Лоясиевича)

В 1963 году Б.Т. Поляк [1] предложил простое условие, достаточное для того, чтобы показать глобальную линейную скорость сходимости градиентного спуска для достаточно гладких задач. Это условие является частным случаем неравенства Лоясиевича, предложенного в том же году [56], и не требует сильной выпуклости (или даже выпуклости). Условие вида (32) называется условием градиентного доминирования, или же условием Поляка–Лоясиевича (PL-условием) [1; 56—58]:

$$f(x) - f(x_*) \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2 \quad \forall x \in \mathbb{R}^n. \tag{32}$$

Как правило, в литературе это условие называют условием Поляка–Лоясиевича, но стоит отметить ещё работу Лежанского [57], в которой оно появилось независимо. По-видимому, правильнее называть (32) условием Лежанского–Поляка–Лоясиевича. Зачастую в теоретических приложениях данное условие используется как ослабление сильной выпуклости, так как любая μ-сильно выпуклая функция удовлетворяет PL-условию с константой μ.

В последние годы условие градиентного доминирования обширно исследуется в различных областях оптимизации и смежных науках. Толчком к возрождению интереса к этому условию послужила работа [58], в которой авторы показали, что PL-условие слабее тех условий, которые ранее использовались для получения линейной скорости сходимости без сильной выпуклости. Также в этой работе PL-условие использовалось для проведения нового анализа рандомизированных и жадных методов покоординатного спуска, а также стохастических градиентных методов в различных постановках. Было предложено и обобщение результатов на проксимальные градиентные методы, позволяющее несложно доказать их линейную скорость сходимости. Попутно были получены результаты сходимости для широкого круга задач машинного обучения: метода наименьших квадратов, логистической регрессии, бустинга, устойчивого метода обратного распространения ошибки, L1-регуляризации, метода опорных векторов.

Приведем несколько примеров задач, в которых условие градиентного доминирования возникает естественным образом.

Пример 9 (PL-условие для нелинейных задач)**.** *Пусть* $f(x) = \frac{1}{2}\|\Phi(x) - y\|_2^2$, *где* $\Phi : \mathbb{R}^d \to \mathbb{R}^n$ *дифференцируемо. Тогда* f *удовлетворяет PL-условию, если существует такое* $\mu > 0$, *что для всех* $x \in \mathbb{R}^d$ *имеет место равномерная невырожденность матрицы Якоби:*

$$\lambda^{\min}\left(\frac{\partial \Phi(x)}{\partial x} \cdot \left[\frac{\partial \Phi(x)}{\partial x}\right]^T\right) \geq \mu, \tag{33}$$

где

$$\frac{\partial \Phi(x)}{\partial x} = \left(\frac{\partial \Phi_i(x)}{\partial x_j}\right)_{i,j=1}^{n,d} = D\Phi(x).$$

Действительно, достаточно написать

$$\|\nabla f(x)\|_2^2 = \|D\Phi(x)^T(\Phi(x) - y)\|_2^2 \geq \mu\|\Phi(x) - y\|_2^2 = 2\mu f(x) \geq 2\mu(f(x) - f(x_*)).$$

В качестве $\Phi(x)$ *может выступать, например, система нелинейных уравнений. В частности, из этого можно получить, что функции Розенброка и Нестерова–Скокова локально удовлетворяют условию Поляка–Лоясиевича.*

Обратим внимание, что допущение (33) предполагает $d \geq n$. Это верно, если, например, Φ задаёт перепараметризованную нейронную сеть. Более точные рассуждения с опорой на использование структуры нейронной сети см. в [59].

В некоторых задачах целевая функция обладает PL-условием не на всём пространстве, а только на каком-то множестве. Однако если траектория вычислительного метода не выводит итеративную последовательность за рамки этого множества, то, конечно, можно применять PL-условие для анализа сходимости метода.

Пример 10 (Композиция сильно (строго) выпуклой и линейной функций)**.** *В [58] показано, что если функция f имеет вид $f(x) = g(Ax)$, где g — сильно выпукла, то f удовлетворяет PL-условию. Функции данного вида часто возникают в машинном обучении (например, метод наименьших квадратов). Если ослабить условие сильной выпуклости функции g до строгой выпуклости, то функция f также будет удовлетворять PL-условию, однако уже не на всем пространстве, а только на произвольном компакте. Например, из данного факта следует, что целевая функция задачи логистической регрессии*

$$f(x) = \sum_{i=1}^{n} \log\left(1 + \exp(b_i a_i^T x)\right)$$

локально (т.е. на всяком компакте) обладает PL-условием.

Также оказывается, что PL-условие (или его различные варианты) оказывается полезным в таких областях как теория оптимального управления [60] и обучение с подкреплением [61; 62], толчком в развитии приведённых работ в которых послужила работа [63]. Приведём пример функции с условием градиентного доминирования, связанной с теорией оптимального управления. Этот достаточно интересный класс функций с условием градиентного доминирования был получен совсем недавно И.Ф. Фатхуллиным и Б.Т. Поляком в [60].

Пример 11 ([60])**.** *Пусть задана линейная система управления*

$$\dot{x}(t) = Ax(t) + Bu(t),$$

где $x(t) \in \mathbb{R}^n$ — состояние системы, а $u(t) \in \mathbb{R}^p$ — управление; с начальными условиями $x(0)$, распределенными случайно с нулевым средним и ковариационной матрицей Σ ($\mathbb{E}x(0)x(0)^T = \Sigma$); с критерием качества вида

$$f(K) = \mathbb{E}\int_0^\infty \left(x(t)^T Q x(t) + u(t)^T R u(t)\right) dt, \quad Q, R \succ 0,$$

заданным на множестве S, с целью найти матрицу обратной связи по состоянию $u(t) = -Kx(t)$, минимизирующую $f(K)$.

Тогда если известен K_0 — стабилизирующий регулятор, то градиент $\nabla f(K)$ удовлетворяет условию Липшица на множестве уровня $S_0 = \{K : f(K) \leq f(K_0)\}$, а функция $f(K)$ удовлетворяет условию градиентного доминирования со следующей константой μ:

$$\mu = \frac{\lambda_1(R)\lambda_1^2(\Sigma)\lambda_1(Q)}{8f(K_*)\left(\|A\|_2 + \frac{\|B\|_2^2 f(K_0)}{\lambda_1(\Sigma)\lambda_1(R)}\right)^2},$$

где K_* — точка минимума $f(K)$ на множестве S. Здесь $\lambda_i(X)$ — собственные числа произвольной квадратной матрицы $X \in \mathbb{R}^{m\times m}$, занумерованные в порядке возрастания их действительных частей, т.е. $Re\lambda_1(X) \leq Re\lambda_2(X) \leq \ldots \leq Re\lambda_m(X)$.

Отметим, что функция $f(K)$ достаточно гладкая (имеет липшицев градиент), но невыпуклая.

Таким образом, условие Поляка–Лоясиевича верно для достаточно большого класса невыпуклых задач. Тем не менее, оно позволяет обосновать сходимость градиентного метода со скоростью геометрической прогрессии. Приведём оценку скорости сходимости градиентного метода для L-гладких функций с условием Поляка–Лоясиевича. Будем рассматривать задачи минимизации функции f, удовлетворяющей PL-условию (32), а также условию Липшица градиента

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall\, x,\, y \in \mathbb{R}^n. \tag{34}$$

Справедлива следующая

Теорема 6. *Пусть дана задача безусловной оптимизации* $\arg\min f(x)$*, где функция f является L-гладкой и удовлетворяет PL-условию с константой μ. Тогда градиентный метод с постоянным шагом $\frac{1}{L}$ вида*

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k) \tag{35}$$

имеет линейную скорость сходимости, то есть

$$f(x^k) - f(x_*) \leq \left(1 - \frac{\mu}{L}\right)^k (f(x^0) - f(x_*)) \leq \exp\left(-\frac{\mu}{L}k\right)(f(x^0) - f(x_*)). \tag{36}$$

Доказательство. В силу L-гладкости функции f

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|_2^2.$$

Пользуясь правилом обновления (35), получаем

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|_2^2. \tag{37}$$

Из PL-условия имеем

$$f(x^k) - f(x_*) \leq \frac{1}{2\mu}\|\nabla f(x^k)\|_2^2. \tag{38}$$

Из (37) и (38) путем преобразований следует

$$f(x^{k+1}) - f(x_*) \leq \left(1 - \frac{\mu}{L}\right)(f(x^k) - f(x_*)).$$

Рекурсивное применение последнего неравенства и дает требуемый результат (36). □

Доказанная верхняя оценка скорости сходимости из теоремы 6 известна уже около 60 лет. До недавнего времени вопрос об её оптимальности оставался открытым. В последние дни 2022 года был выложен препринт [64], в котором обоснована оптимальность этой оценки на классе гладких задач с условием Поляка–Лоясиевича.

Отметим, что существует аналог условия Поляка–Лоясиевича для седловых задач, так называемое двухстороннее условие Поляка–Лоясиевича, которое позволяет получить ряд схожих результатов [65—67]. Будем рассматривать седловую задачу $\min\limits_x \max\limits_y f(x,y)$. Говорят, что функция $f(x,y)$ удовлетворяет двустороннему PL-условию, если существуют такие константы μ_1 и μ_2, что $\forall x, y$:

$$\begin{cases} \|\nabla_x f(x,y)\|_2^2 \geq 2\mu_1 \left(f(x,y) - \min\limits_x f(x,y) \right); \\ \|\nabla_y f(x,y)\|_2^2 \geq 2\mu_2 \left(\max\limits_y f(x,y) - f(x,y) \right). \end{cases}$$

Примеры функций, удовлетворяющих двустороннему PL-условию:

- $f(x,y) = F(Ax, By)$, где $F(\cdot,\cdot)$ — сильно-выпукло-сильно-вогнутая и A, B — произвольные матрицы.
- Невыпукло-невогнутая $f(x,y) = x^2 + 3\sin^2 x \sin^2 y - 4y^2 - 10\sin^2 y$ ($\mu_1 = 1/16$, $\mu_2 = 1/14$) [65].
- Релаксация задачи Robust Least Squares (RLS):

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} F(x,y);$$
$$F(x,y) := \|Ax - y\|_M^2 - \lambda \|y - y_0\|_M^2,$$

  где M — положительно полуопределена и $\|x\|_M^2 = x^T M x$. RLS минимизирует невязку наихудшего случая при заданном ограниченном детерминированном возмущении δ на зашумленном векторе измерений $y_0 \in \mathbb{R}^m$ и матрице коэффициентов $A \in \mathbb{R}^{m \times n}$. Саму задачу RLS можно сформулировать так ($\rho > 0$) [68]:

$$\min_{x \in \mathbb{R}^n} \max_{\delta : \|\delta\|_2 \leq \rho} \|Ax - y\|_2^2, \text{ где } \delta = y_0 - y \in \mathbb{R}^m.$$

  Известно, что при $\lambda > 1$ $F(x,y)$ удовлетворяет двустороннему PL-условию, поскольку $F(x,y)$ можно представить как комбинацию аффинной функции и сильно-выпукло-сильно-вогнутой функции.

Для седловых задач с двусторонним условием градиентного доминирования может использоваться вариация градиентного метода — метод градиентного спуска-подъема. Этот подход сейчас довольно популярен [65; 66; 69], но был подробно проанализирован ещё в малоизвестной статье Бакушинского–Поляка [70].

Далее, поговорим о поведении градиентного метода (сходимость и возможные правила ранней остановки) для класса достаточно гладких задач с условием градиентного доминирования, когда градиент целевой функции доступен с некоторой погрешностью. Б.Т. Поляк считал важными вопросы исследования влияния погрешностей доступной информации на гарантии сходимости численных методов, занимался ими и уделил им немало внимания в своей книге [2]. Так, этой тематике посвящена одна из его последних работ [71], о которой мы немного расскажем.

Сфокусируемся на двух основных известных типах неточной информации о градиенте: абсолютная и относительная погрешности. Типичными примерами областей, в которых возникает данная проблема неточного градиента, являются безградиентная оптимизация, в которой используется приближенно вычисляемый градиент [72—74], а также

оптимизация в бесконечномерных пространствах, связанная с обратными задачами [75; 76].

Начнем с вопроса влияния на качество выдаваемого решения абсолютной погрешности в градиенте. Будем рассматривать задачу минимизации функции f, удовлетворяющей PL-условию (32), а также условию Липшица градиента (34). Считаем, что методу доступно не точное, а приближённое значение градиента $\widetilde{\nabla} f(x)$ в любой запрашиваемой точке x:

$$\nabla f(x) = \widetilde{\nabla} f(x) + v(x), \quad \text{причём} \quad \|v(x)\|_2 \leq \Delta$$

при некотором фиксированном $\Delta > 0$. Тогда (32) означает, что

$$f(x) - f(x_*) \leq \frac{1}{\mu}(\|\widetilde{\nabla} f(x)\|_2^2 + \Delta^2) \quad \forall\, x \in \mathbb{R}^n, \tag{39}$$

откуда

$$\|\widetilde{\nabla} f(x)\|_2^2 \geq \mu(f(x) - f(x_*)) - \Delta^2 \quad \forall\, x \in \mathbb{R}^n.$$

Заметим, что вопросы исследования влияния погрешностей градиента на оценки скорости сходимости методов первого порядка ужа достаточно давно привлекают многих исследователей (см., например, [2; 77—80]). Однако мы сфокусируемся именно на выделенном классе, вообще говоря, невыпуклых задач. Невыпуклость целевой функции задачи, а также использование на итерациях неточно заданного градиента могут приводить к разным проблемам. В частности, при отсутствии каких-либо правил ранней остановки возможно достаточно большое удаление траектории градиентного метода от начальной точки. Это может быть проблемным, если начальное положение метода уже обладает определёнными хорошими свойствами. С другой стороны, неограниченное удаление траектории градиентного метода в случае градиентного метода с помехами может привести к удалению от искомого точного решения. Опишем некоторые ситуации такого типа.

В качестве простого примера не сильно выпуклой функции, удовлетворяющей условию градиентного доминирования, можно рассмотреть

$$f(x) = \langle Ax, x\rangle, \tag{40}$$

где $A = \mathrm{diag}(L, \mu, 0)$ — диагональная матрица 3-го порядка с ровно двумя положительными элементами $L > \mu > 0$. Если для задачи минимизации функции (40) допустить наличие погрешности градиента $v(x) = (0, 0, \Delta)$ при $\Delta > 0$, то при $x^0 = (0, 0, 0)$ и $h_k > 0$ последовательность $x^{k+1} = x^k - h_k \widetilde{\nabla} f(x^k)$ стремится к бесконечности при неограниченном увеличении k. Далее, можно рассмотреть функцию Розенброка двух переменных $x = (x_{(1)}, x_{(2)})$

$$f(x) = 100\left(x_{(2)} - \left(x_{(1)}\right)^2\right)^2 + \left(1 - x_{(1)}\right)^2.$$

При $x^0 = (1, 1) = x_*$ на каждом шаге градиентного метода возможна такая погрешность градиента $v(x^k)$, что $x^k_{(2)} = \left(x^k_{(1)}\right)^2$ и без остановки траектория может уходить весьма далеко от точного решения x_*. Аналогично траектория градиентного метода может уходить к бесконечности для целевой функции двух переменных $f(x) = (x_{(2)} - (x_{(1)})^2)^2$.

Немного расскажем об одной из последних работ, подготовленных Б.Т. Поляком с соавторами [71]. В этой статье поставлена цель исследовать оценку расстояния $\|x^N - x^0\|_2$

для выдаваемых градиентным методом точек x^N и предложить правило ранней остановки, которое гарантирует некоторый компромисс между стремлением достичь приемлемого качества выдаваемого методом решения задачи по функции и не столь существенным удалением траектории от выбранной начальной точки метода. Отметим, что правила ранней остановки в итеративных процедурах активно исследуются для самых разных типов задач. По-видимому, впервые идеология раннего прерывания итераций была предложена в статье [81], посвящённой методике приближённого решения возникающих при регуляризации некорректных или плохо обусловленных задач (в упомянутой работе рассматривалась задача решения линейного уравнения). Ранняя остановка в этом случае нацелена на преодоление проблемы потенциального накопления погрешностей регуляризации исходной задачи. К данной тематике примыкают известные подходы, связанные с ранней остановкой методов первого порядка в случае использования на итерациях неточной информации о градиенте (см. [2], гл. 6, параграф 1, а также, к примеру, недавний препринт [80]). Однако известные прежде результаты для выпуклых (не сильно выпуклых) задач отличаются по сравнению с полученным в рассматриваемой работе тем, что обычно гарантируется достижение худшего уровня по функции (если сравнить с комментарием после теоремы 2 параграфа 1 гл. 6 из [2]) или оценки вида $\|x^N - x_*\|_2 \leq \|x^0 - x_*\|_2$ без исследования $\|x^N - x^0\|_2$, где $\{x^k\}_{k\in\mathbb{N}}$ — образуемая методом последовательность, x_* — ближайшее к точке старта метода x^0 точное решение задачи минимизации f.

В предположении доступности значений параметров $L > 0$ и $\Delta > 0$ применим к задаче минимизации f градиентный метод вида

$$x^{k+1} = x^k - \frac{1}{L}\widetilde{\nabla} f(x^k). \tag{41}$$

Тогда ввиду (34) для метода (41) имеем следующие соотношения:

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 =$$

$$= f(x^k) + \frac{1}{2L}\|\nabla f(x^k) - \widetilde{\nabla} f(x^k)\|_2^2 - \frac{\|\nabla f(x^k)\|_2^2}{2L} \leq$$

$$\leq f(x^k) + \frac{\Delta^2}{2L} - \frac{1}{2L}\|\nabla f(x^k)\|_2^2,$$

т.е.

$$f(x^{k+1}) - f(x^k) \leq \frac{\Delta^2}{2L} - \frac{1}{2L}\|\nabla f(x^k)\|_2^2. \tag{42}$$

Суммирование неравенств (42) приводит нас к оценке

$$\min_{k=0,N-1} \|\nabla f(x^k)\|_2 \leq \sqrt{\Delta^2 + \frac{2L(f(x^0) - f(x_*))}{N}} \leq \Delta + \sqrt{\frac{2L(f(x^0) - f(x_*))}{N}}, \tag{43}$$

что указывает на потенциальную возможность расходимости градиентного метода с абсолютной погрешностью задания градиента. Конкретные примеры таких ситуаций описаны выше.

С учётом (32) тогда получаем, что

$$f(x^{k+1}) - f(x^k) \leq \frac{\Delta^2}{2L} - \frac{2\mu(f(x^k) - f(x_*))}{2L} = -\frac{\mu}{L}(f(x^k) - f(x_*)) + \frac{\Delta^2}{2L},$$

откуда

$$f(x^{k+1}) - f(x_*) \leq \left(1 - \frac{\mu}{L}\right)(f(x^k) - f(x_*)) + \frac{\Delta^2}{2L} \leq$$

$$\leq \left(1 - \frac{\mu}{L}\right)^{k+1}(f(x^0) - f(x_*)) + \frac{\Delta^2}{2L}\left(1 + 1 - \frac{\mu}{L} + \cdots + \left(1 - \frac{\mu}{L}\right)^k\right) <$$

$$< \left(1 - \frac{\mu}{L}\right)^{k+1}(f(x^0) - f(x_*)) + \frac{\Delta^2}{2\mu},$$

т.е.

$$f(x^{k+1}) - f(x_*) \leq \left(1 - \frac{\mu}{L}\right)^{k+1}(f(x^0) - f(x_*)) + \frac{\Delta^2}{2\mu}. \tag{44}$$

Оценки (43) и (44), вообще говоря, неулучшаемы. К примеру, известны нижние оценки уровня точности по функции $O\left(\frac{\Delta^2}{2\mu}\right)$ для градиентного метода с абсолютной погрешностью задания градиента (см., например, раздел 2.11.1 пособия [37], а также имеющиеся там ссылки) даже на классе сильно выпуклых функций.

В [71] для градиентного метода с постоянным шагом при достаточно малом значении возмущённого градиента получена теорема, в которой указан уровень точности по функции, который возможно гарантировать после выполнения предложенного правила ранней остановки. Данный результат можно применять ко всяким L-гладким невыпуклым задачам. Далее, с использованием PL-условия получено уточнение этого результата — достаточное количество итераций градиентного метода для достижения желаемого качества точки выхода $\widehat{x}$ по функции $f(\widehat{x}) - f(x_*) = O\left(\frac{\Delta^2}{\mu}\right)$.

Теперь перейдем к ситуации относительной неточности градиента. Будем рассматривать поведения методов градиентного типа для выделенного класса достаточно гладких задач с условием Поляка–Лоясиевича в предположении доступности для метода в каждой текущей точке градиента с относительной погрешностью, т.е.

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \alpha \|\nabla f(x)\|_2, \tag{45}$$

где под $\tilde{\nabla} f(x)$, как и ранее, мы понимаем неточный градиент, а $\alpha \in [0, 1)$ — некоторая константа, указывающая на величину относительной погрешности градиента. Данное условие на неточный градиент было введено и изучено в работах [2; 82].

Из (45) можно получить вариант условия градиентного доминирования (32) для относительно неточного градиента

$$f(x) - f(x_*) \leq \frac{1}{2\mu(1-\alpha)^2}\|\tilde{\nabla} f(x)\|_2^2.$$

Описанную задачу предлагается так же решать методом градиентного типа вида

$$x^{k+1} = x^k - h_k \tilde{\nabla} f(x^k). \tag{46}$$

Оказывается, что можно подобрать так параметры шагов h_k, чтобы гарантировать сохранение результата о сходимости метода со скоростью геометрической прогрессии в случае относительной неточности градиента (хоть и с большим знаменателем).

Если при реализации метода (46) градиент доступен с известной относительной погрешностью $\alpha \in [0, 1)$ и известен параметр $L > 0$, то выбор шага $h_k = \frac{1}{L}\frac{(1-\alpha)}{(1+\alpha)^2}$ приводит

к следующим результатам (см. параграф 1 из пособия [83] и имеющиеся там ссылки).

$$
\begin{aligned}
f(x^{k+1}) - f(x^k) &\leq \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 \\
&= -h_k \langle \nabla f(x^k), \tilde{\nabla} f(x^k) \rangle + \frac{L h_k^2}{2}\|\tilde{\nabla} f(x^k)\|_2^2 \\
&= -h_k \|\nabla f(x^k)\|_2^2 - h_k \langle \nabla f(x^k), \tilde{\nabla} f(x^k) - \nabla f(x^k) \rangle + \frac{L h_k^2}{2}\|\tilde{\nabla} f(x^k)\|_2^2 \\
&\leq -h_k \|\nabla f(x^k)\|_2^2 + h_k \|\nabla f(x^k)\|_2 \|\tilde{\nabla} f(x^k) - \nabla f(x^k)\|_2 + \frac{L h_k^2}{2}\|\tilde{\nabla} f(x^k)\|_2^2 \\
&\leq -h_k \|\nabla f(x^k)\|_2^2 + \alpha h_k \|\nabla f(x^k)\|_2^2 + (1+\alpha)^2 \frac{L h_k^2}{2}\|\nabla f(x^k)\|_2^2 \\
&= \left(-(1-\alpha) h_k + (1+\alpha)^2 \frac{L h_k^2}{2} \right) \|\nabla f(x^k)\|_2^2.
\end{aligned}
$$

После подстановки h_k (которое, заметим, минимизирует выражение в скобках в последнем выражении) получаем

$$
f(x^{k+1}) - f(x^k) \leq -\frac{1}{2L} \frac{(1-\alpha)^2}{(1+\alpha)^2} \left\| \nabla f(x^k) \right\|_2^2. \tag{47}
$$

Пользуясь условием Поляка–Лоясиевича (32), имеем следующую оценку на скорость сходимости по функции

$$
f(x^{k+1}) - f(x_*) \leq \left(1 - \frac{\mu}{L} \frac{(1-\alpha)^2}{(1+\alpha)^2} \right) \left(f(x^k) - f(x_*) \right),
$$

т.е.

$$
f(x^N) - f(x_*) \leq \left(1 - \frac{\mu}{L} \frac{(1-\alpha)^2}{(1+\alpha)^2} \right)^N \left(f(x^0) - f(x_*) \right).
$$

Если вместо PL-условия (32) предполагать выполнение условия сильной выпуклости, то приведенную оценку можно уточнить [84]:

$$
\frac{(1-\alpha)^2}{(1+\alpha)^2} \to O\left(\frac{1-\alpha}{1+\alpha} \right).
$$

В работе [85] предлагается адаптивный алгоритм для случая относительной погрешности градиента. По сути, предлагается критерий выхода из итерации, содержащий норму неточного градиента $\|\tilde{\nabla} f(x^k)\|_2$. Это обстоятельство затрудняет использование подхода с оценками типа (47) для нормы точного градиента. Поэтому рассматривается альтернативный вариант выбора шага для метода (46) с относительной погрешностью задания градиента, который позволит получить приемлемый аналог оценки (47) для квадрата нормы неточного градиента. Аналогично ранее рассмотренным случаям, пользуясь L-липшицевостью градиента и условием (45), можно вывести следующее неравенство [85]

$$
\begin{aligned}
f(x^{k+1}) \leq f(x^k) + \left\langle \tilde{\nabla} f(x^k), x^{k+1} - x^k \right\rangle + \frac{L}{2} \left\| x^{k+1} - x^k \right\|_2^2 + \\
+ \frac{\alpha}{1-\alpha} \left\| \tilde{\nabla} f(x^k) \right\|_2 \left\| x^{k+1} - x^k \right\|_2.
\end{aligned} \tag{48}
$$

Таким образом, при $\alpha \in [0; 0.5)$ для адаптивного варианта градиентного метода (46) с

$$h_k = \frac{1}{L_{k+1}} \frac{1-2\alpha}{1-\alpha} \tag{49}$$

и критерием выхода из итерации (до его выполнения L_{k+1} увеличивается в 2 раза)

$$f(x^{k+1}) \leq f(x^k) + \left\langle \tilde{\nabla} f(x^k), x^{k+1} - x^k \right\rangle + \frac{L_{k+1}}{2} \left\| x^{k+1} - x^k \right\|_2^2 + \\ + \frac{\alpha}{1-\alpha} \left\| \tilde{\nabla} f(x^k) \right\|_2 \left\| x^{k+1} - x^k \right\|_2 \tag{50}$$

неравенство (48) гарантирует выход из итерации при $L_{k+1} \geq L$. Тогда оценка скорости сходимости метода (46) с шагом (49) имеет вид

$$f(x^N) - f(x_*) \leq \prod_{k=0}^{N-1} \left(1 - \frac{\mu}{L_{k+1}} (1-2\alpha)^2 \right) (f(x^0) - f(x_*)). \tag{51}$$

Если $L_0 \leq 2L$, то последнее неравенство означает сходимость рассматриваемого метода со скоростью геометрической прогрессии

$$f(x^N) - f(x_*) \leq \left(1 - \frac{\mu}{2L} (1-2\alpha)^2 \right)^N \left(f(x^0) - f(x_*) \right).$$

Отметим, что оценка (51) может быть применима в случае неизвестного значения L. Кроме того, даже в случае известной оценки на L потенциально возможно улучшение качества точки выхода метода при условии $L_{k+1} < L$.

3.2. Метод условного градиента (Франк–Вульфа–Левитина–Поляка)

В этом подпункте мы вспомним о направлении научных исследований Б.Т. Поляка, связанном с методами условного градиента. К примеру, таким методам посвящена достаточно известная работа [10]. Несмотря на то, что методы условного градиента по числу итераций проигрывают оптимальным на классе выпуклых гладких задач ускоренным методам (см. раздел 3.3) для решения задач со структурой (например, на единичном симплексе или прямом произведении таких симплексов), по времени работы эти методы могут быть существенно эффективнее ускоренных из-за возможной дешевизны итерации [86—89]. На данный момент методы условного градиента достаточно хорошо разработаны, см., например, обзорную статью [90] и монографию [91]. В частности, среди последних достижений можно отметить безградиентные варианты метода условного градиента для задач стохастической оптимизации [92], а также вариант метода условного градиента для децентрализованных оптимизационных задач на меняющихся графах [93].

Рассмотрим задачу выпуклой оптимизации ($f(x)$ – выпуклая функция, Q – ограниченное выпуклое множество):

$$\min_{x \in Q} f(x). \tag{52}$$

Обозначим решение задачи (52) через x_* (если решение не единственно, то x_* – какое-то из решений).

Алгоритм 3 Метод условного градиента [10; 91]

Require: f — выпуклая, непрерывно дифференцируемая функция; Q — допустимое множество, выпуклое и компактное; x^1 — начальная точка; N — количество итераций.

Ensure: точка x^N

1: **for** $k = 1, \ldots, N-1$ **do**
2: $\quad \gamma_k = \frac{2}{k+1}$, $0 \le \gamma_k \le 1$
3: $\quad y^k = \arg\min_{y \in Q} \langle \nabla f(x^k), y \rangle$
4: $\quad x^{k+1} = (1-\gamma_k)x^k + \gamma_k y^k$
5: **end for**
6: **return** x^N

Теорема 7 (см. [10; 91]). *Пусть $\nabla f(x)$ удовлетворяет на Q условию Липшица с константой L по отношению к некоторой норме $\|\cdot\|$: для всех $x, y \in Q$ выполняется*

$$\|\nabla f(y) - \nabla f(x)\|_* \le L\|y - x\|,$$

$R = \sup_{x,\, y \in Q} \|x - y\|$, $\gamma_k = \frac{2}{k+1}$ для $k \ge 1$. Тогда для любого $N \ge 2$ выполняется

$$f(x^N) - f(x_*) \le \frac{2LR^2}{N+2}. \tag{53}$$

Доказательство. Из условия Липшица на градиент $f(x)$ и выпуклости $f(x)$ имеем оценку

$$0 \le f(x) - f(y) - \langle \nabla f(y), x - y \rangle \le \frac{L}{2}\|x - y\|^2$$

для всех $x, y \in Q$. Получаем цепочку неравенств:

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\le \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ &= \gamma_k \langle \nabla f(x^k), y^k - x^k \rangle + \frac{L\gamma_k^2}{2}\|y^k - x^k\|^2 \\ &\le \gamma_k \langle \nabla f(x^k), x_* - x^k \rangle + \frac{L\gamma_k^2}{2}R^2 \le \gamma_k (f(x_*) - f(x^k)) + \gamma_k^2 \frac{LR^2}{2}. \end{aligned}$$

Введём обозначение $\delta_k = \frac{f(x^k) - f(x_*)}{LR^2}$. Тогда неравенство можно переписать в виде

$$\delta_{k+1} \le (1 - \gamma_k)\delta_k + \frac{\gamma_k^2}{2} = \left(1 - \frac{2}{k+1}\right)\delta_k + \frac{2}{(k+1)^2}.$$

Начиная с неравенства

$$\delta_2 \le \left(1 - \frac{2}{1+1}\right)\delta_1 + \frac{2}{(1+1)^2} = \frac{1}{2},$$

применением индукции по k получаем желаемый результат. □

Как видим из предыдущего результата, при получении оценки скорости сходимости метода условного градиента не важен выбор конкретной нормы. Важно, чтобы параметры L и R оценивались относительно согласованных норм $\|\cdot\|$ и $\|\cdot\|_*$.

Следуя [2], покажем, что оценка (53) не может быть улучшена (с точностью до числового множителя). Для этого выберем $f(x_1, x_2) = x_1^2 + (1+x_2)^2$, $Q = \{x : \ |x_1| \le 1, 0 \le x_2 \le 1\}$. Тогда $x_* = (0,0)^T$. При этом $y^k = (1,0)^T$ при $y_1^k < 0$ и $y^k = (-1,0)^T$ при $y_1^k > 0$. Несложно показать, что для этого примера $f(x^N) - f(x_*) \simeq \frac{4}{N}$. Причем эта ситуация типична для случая, когда Q — многогранник, а минимум достигается в одной из его вершин, поскольку в качестве y^k могут выступать лишь вершины Q.

Отметим также, что в рассматриваемой работе Левитина–Поляка [10] впервые для метода условного градиента появилась идея использовать параметр шага, зависящий от константы Липшица градиента целевой функции L. В современных работах [90; 91; 94] такой шаг обычно применяют в виде

$$\gamma_k = \gamma_k(L) := \min\left\{-\frac{\nabla f(x^k)^T(y^k - x^k)}{L\|y^k - x^k\|^2}, 1\right\}, \tag{54}$$

где L — константа Липшица ∇f. Размер соответствующего шага может рассматриваться как результат минимизации квадратичной модели $m_k(\cdot\,; L)$, переоценивающей f вдоль прямой $x^k + \gamma_k(y^k - x^k)$,

$$m_k(\gamma_k; L) = f(x^k) + \gamma_k \nabla f(x^k)^T(y^k - x^k) + \frac{L\gamma_k^2}{2}\|y^k - x^k\|^2 \ge f(x^k + \gamma_k(y^k - x^k)), \tag{55}$$

где последнее неравенство следует из стандартной леммы о спуске. Интерес такого выбора шага для метода Франк–Вульфа заключается, в частности, в возможности предложить достаточные условия убывания на итерации невязки по функции не менее чем в 2 раза. Точнее говоря, согласно лемме 2 из [90] для всякой выпуклой функции f в случае, если на k-й итерации в (55) верно $\gamma_k = 1$, то верно неравенство

$$f(x^{k+1}) - f^* \le \frac{1}{2}\min\left(L\|y^k - x^k\|^2, f(x^k) - f^*\right). \tag{56}$$

При введении дополнительного условия на функцию, например градиентного доминирования, для шага вида (54) можно получить и результаты о сходимости метода типа условного градиента со скоростью геометрической прогрессии. Например, известен такой результат для выпуклых функций с условием Поляка–Лоясиевича [90]. При этом такие функции могут не быть сильно выпуклыми (например, целевая функция из задачи логистической регрессии не сильно выпукла, но удовлетворяет условию градиентного доминирования на всяком компакте [58]).

Теорема 8. *Пусть Q — компактное выпуклое множество диаметра R, а f — выпуклая функция, удовлетворяющая условию градиентного доминирования с константой $\mu > 0$. Пусть также существует точка минимума f и она лежит во внутренности множества Q ($x_* \in Int(Q)$), т. е. существует $r > 0$, такой, что $B(x_*, r) \subset Q$. Тогда при любых $k \ge 1$ верно*

$$f(x^k) - f^* \le (f(x^0) - f^*)\prod_{i=1}^{k}\varphi_i,$$

где

$$\varphi_i = \begin{cases} \frac{1}{2}, & \text{если } \alpha_i = 1, \\ 1 - \frac{r^2}{Lc^2R^2}, & \text{если } \alpha_i < 1, \end{cases}$$

и $c^2 = \frac{1}{2\mu}$. Таким образом, имеет место сходимость по функции со скоростью геометрической прогрессии, знаменатель которой в худшем случае равен $\max\left\{\frac{1}{2}, 1 - \frac{r^2}{Lc^2R^2}\right\}$.

Отметим, что недавно в работе [94] доказан аналог теоремы 8 для варианта метода условного градиента с адаптивным подбором шага.

Что касается метода условного градиента, исследованного в самой работе Левитина–Поляка [10], то обратим внимание на две отмеченные там его особенности. Первая из них касается нижней оценки скорости сходимости такого метода. Точнее говоря, в замечании 1 на стр. 805 [10] отмечено, что оценка $O\left(\frac{1}{k}\right)$ невязки по функции неулучшаема с точностью до константы на классе выпуклых гладких функций f. Для того времени рассуждения на тему нижних оценок были редки, поскольку направление исследований численных методов оптимизации, основанное на нижних оценках их эффективности для классов задач, зародилось несколько позднее в монографии [52]. Вторая особенность метода условного градиента, отмеченная впервые в [10], касается повышения скоростных гарантий не путём накладывания дополнительных условий на целевую функцию, а посредством сужения класса допустимых множеств задачи Q. Например, при добавлении требования, что множество Q является сильно выпуклым. Множество Q называется сильно выпуклым, если существует функция $\delta(\tau) = \gamma\tau^2$ $(\gamma > 0)$ такая, что $(x+y)/2 + z \in Q$ для любых $x, y \in Q$ и любого z: $\|z\| \leq \delta\left(\|x-y\|\right)$. Существуют эквивалентные определения сильной выпуклости множества. Например, как было показано в [95] (теорема 1 на стр. 190), множество является сильно выпуклым тогда и только тогда, когда оно представимо в виде пересечения шаров одного радиуса. Более подробно см. на стр. 789 [10], а также в [95—97], метод сходится по аргументу со скоростью геометрической прогрессии при условии, что $\|\nabla f(x)\| \geq \varepsilon > 0$ на Q.

Исследование разных вариантов метода условного градиента продолжается и в настоящее время [90; 91; 94; 98]. В работе [98] имеется обзор результатов последних лет по оценкам сложности для метода Франк–Вульфа, предложен универсальный (адаптивный по параметру гладкости задачи) вариант метода для задач с равномерно выпуклой структурой. В [94] адаптивный вариант метода Франк–Вульфа с шагом (54) исследуется для выпуклых задач (вообще говоря, без равномерной структуры), причём детально проработан соответствующий аналог оценки (56).

3.3. Ускоренные методы выпуклой оптимизации

Для наглядности изложения рассмотрим задачу безусловной выпуклой оптимизации

$$\min_x f(x), \tag{57}$$

в которой выполняется условие L-Липшицевости градиента в 2-норме, см. (34), а x – принадлежит гильбертову пространству. Также для наглядности будем опускать числовые множители в оценках скорости сходимости.

В кандидатской диссертации Б.Т. Поляка в 1963 году (см. также [8]) было показано, что градиентный спуск на такой задаче

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k)$$

будет сходиться следующим образом

$$f(x^N) - f(x_*) \lesssim \frac{LR^2}{N},$$

где $R = \|x^0 - x_*\|_2$ – расстояние от точки старта до ближайшего (в 2-норме) к точке старта решения x_* задачи (57) (не ограничивая общности, везде в разделе 3.3 можно считать,

что все константы, типа L и μ, определены на шаре с центром в точке x^0 и радиусом $2R$ [51; 83]). Оптимальная ли эта оценка? Оказывается, что даже если выбирать шаг метода $h = 1/L$ как-то по-другому, улучшить такую оценку на рассматриваемом классе задач можно только на числовой множитель порядка 4 [99]. Однако это совсем не означает, что если рассмотреть какой-то другой метод градиентного типа, то на нем также нельзя будет получить существенно лучшую оценку скорости сходимости. Действительно, как минимум на квадратичных задачах выпуклой оптимизации метод сопряженных градиентов (первая итерация совпадает с итерацией типа градиентного спуска, аналогично и для приводимых далее моментных методов):

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1}), \tag{58}$$

где

$$(\alpha_k, \beta_k) \in \operatorname{Arg}\min_{\alpha,\beta} f\left(x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1})\right)$$

дает существенно лучшую оценку скорости сходимости (при дополнительных предположениях о распределении спектра матрицы квадратичной формы оценка может быть дополнительно улучшена [100]):

$$f(x^N) - f(x_*) \lesssim \frac{LR^2}{N^2}. \tag{59}$$

Наиболее тонкое исследование скорости сходимости метода сопряженных градиентов (58) (в том числе в условиях неточностей) имеется в работе [101]. Стоимость итерации такого метода (в виду возможности решить задачу поиска α_k и β_k аналитически для квадратичных задач) будет по порядку такой же, как стоимость итерации градиентного метода. Аналогичные оценки можно написать и для μ-сильно выпуклых задач:

$$f(x^N) - f(x_*) \lesssim LR^2 \exp\left(-\frac{\mu}{L} N\right) \tag{60}$$

для градиентного спуска и:

$$f(x^N) - f(x_*) \lesssim LR^2 \exp\left(-2\sqrt{\frac{\mu}{L}} N\right) \tag{61}$$

для метода сопряженных градиентов (58) на квадратичной задаче.

Заметим, что одна из форм записи метода сопряженных градиентов (58) в случае μ-сильно выпуклой целевой функции приводит к методу Чебышёва [102]:

$$x^{k+1} = x^k - \frac{4\delta_k}{L-\mu}\nabla f(x^k) + \left(\frac{2\delta_k\,(L+\mu)}{L-\mu} - 1\right)\left(x^k - x^{k-1}\right), \tag{62}$$

$$x^1 = x^0 - \frac{2}{L+\mu}\nabla f(x^0),$$

$$\delta_{k+1} = \frac{1}{2\frac{L+\mu}{L-\mu} - \delta_k}, \quad \delta_1 = \frac{1}{2\frac{L+\mu}{L-\mu} + 1}.$$

Метод тяжёлого шарика Поляка (или импульсный / моментный метод Поляка) [9] при этом выглядит так:

$$x^{k+1} = x^k - \frac{4}{\left(\sqrt{L} + \sqrt{\mu}\right)^2}\nabla f(x^k) + \frac{\left(\sqrt{L} - \sqrt{\mu}\right)^2}{\left(\sqrt{L} + \sqrt{\mu}\right)^2}\left(x^k - x^{k-1}\right). \tag{63}$$

Этот метод получается в асимптотике $k \to \infty$ из метода Чебышёва (62), поскольку

$$\delta_\infty = \frac{1}{2\frac{L-\mu}{L+\mu} - \delta_\infty},$$

следовательно,

$$\delta_\infty = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}.$$

Приведенный вывод (с учетом оптимальности метода Чебышёва на квадратичных задачах) дает надежду, что метод тяжелого шарика (63) в некотором смысле асимптотически оптимальный. Оказывается, что «в среднем» так оно и есть [103].

Исходно метод тяжелого шарика (63) был получен с помощью первого метода Ляпунова [2], который сводит анализ скорости локальной сходимости метода на классе выпуклых задач к анализу скорости глобальной сходимости для квадратичных выпуклых задач. Б.Т. Поляк подбирал коэффициенты α и β постоянными. Подобранные коэффициенты гарантировали локальную скорость сходимости метода (63) аналогичную скорости сходимости метода сопряженных градиентов (61). Отметим, что при этом имеется некоторая физическая аналогия в полученном таким образом методе (63) с овражным методом Гельфанда–Цетлина [104]. Однако как показал последующий анализ метода тяжелого шарика (63), в общем случае нельзя гарантировать его глобальную сходимости [105]. Ее можно добиться за счет некоторых дополнительных предположений (о гладкости), но скорость глобальной сходимости получалась уже не лучше, чем у обычного градиентного спуска [106], причем то что лучше и не получится удалось доказать [107]. Также на примере тяжелого шарика (63) хорошо демонстрируется общая особенность ускоренных методов – немонотонное убывание целевой функции по ходу итерационного процесса и наличие циклов длин $\sim \sqrt{L/\mu}$ (после каждого цикла невязка по функции убывает в ~ 2 раза). Изящное объяснение этому явлению можно найти в работах [108; 109].

Метод тяжелого шарика (63) сыграл очень важную роль в развитии ускоренных методов выпуклой оптимизации. Общее направление исследований тут можно описать как попытку сделать такую версию метода сопряженных градиентов, для которой бы удалось доказать глобальную оценку скорости сходимости, аналогичную приведенным выше, для обычного метода сопряженных градиентов на квадратичных задачах. Так родились, например, методы Флетчера–Ривса, Полака–Рибьера–Поляка [12]. Точку здесь удалось поставить А.С. Немировскому в конце 70-х годов прошлого века [52; 110], попутно показав, что полученные оценки скорости сходимости (59) и (61) уже не могут быть дальше улучшены никакими другими методами, использующими градиент целевой функции (улучшить немного можно лишь числовой множитель). То есть получилось, что сложность класса гладких задач (сильно) выпуклой оптимизации с точностью до числовых множителей совпадает со сложностью аналогичных классов задач (сильно) выпуклой квадратичной оптимизации. Отметим, что для распределенной оптимизации (а точнее, федеративного обучения) недавно было обнаружено, что такая аналогия уже не имеет места [111; 112]. Отметим также, что ускоренные методы А.С. Немировского (известно как минимум три таких метода) требовали на каждой итерации решение вспомогательной маломерной задачи оптимизации. Избавиться от этого удалось в кандидатской диссертации Ю.Е. Нестерова (научным руководителем был Б.Т. Поляк) в 1983

году [113]:

$$x^{k+1} = x^k - \frac{1}{L}\nabla f\left(x^k + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\left(x^k - x^{k-1}\right)\right) + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\left(x^k - x^{k-1}\right). \tag{64}$$

Такой ускоренный (моментный) метод Нестерова будет сходиться аналогично (61) уже глобально на классе гладких μ-сильно выпуклых задач (аналогично можно предложить вариант метода и для выпуклых задач $\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}} \to \frac{k-1}{k+2}$ в (64)). Отметим, что недавно были предложены (см. обзор [102]) такие варианты ускоренного метода Нестерова (с аналогичными по порядку трудозатратами на каждой итерации), для которых удалось доказать, что их скорость сходимости в выпуклом и сильно выпуклых случаях оптимальны (без оговорки, «с точностью до числового множителя») на классе градиентных методов выпуклой оптимизации, использующих всевозможные линейные комбинации полученных на предыдущих итерациях градиентов (в том числе, допускается вспомогательная минимизация на подпространствах натянутых на полученные градиенты). Например, в μ-сильно выпуклом случае оптимальный алгоритм выглядит таким образом.

Алгоритм 4 Ускоренный метод Тейлора–Дрори [102]

Require: f — μ-сильно выпуклая функция с L-липшицевым градиентом, точка старта x^0

SET: $z^0 = x^0$, $A_0 = 0$, $q = \mu/L$
for $k = 0, \ldots, N-1$ **do**
  $A_{k+1} = \frac{(1+q)A_k + 2\left(1+\sqrt{(1+A_k)(1+qA_k)}\right)}{(1-q)^2}$
  $\tau_k = 1 - \frac{A_k}{(1-q)A_{k+1}}$ и $\delta_k = \frac{1}{2}\frac{(1-q)^2 A_{k+1} - (1+q)A_k}{1+q+qA_k}$
  $y^k = x^k + \tau_k\left(z^k - x^k\right)$
  $x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$
  $z^{k+1} = (1-q\delta_k)z^k + q\delta_k y^k - \frac{\delta_k}{L}\nabla f(y^k)$
end for
return z^N

Как ни странно, в 1983 году статья Ю.Е. Нестерова не вызвала какого-то большого ажиотажа. О ней вспомнили лишь спустя 20 лет уже в новом столетии, когда размеры новых задач, которые требовалось решать в анализе данных, позволяли использовать только градиентные методы (а не, скажем, методы внутренней точки) или их стохастические варианты. Во многом этому способствовало появление в 2004 году первого издания лекций Ю.Е. Нестерова по выпуклой оптимизации [51], в которых центральную роль как раз и заняло современное (на тот момент) изложение ускоренных методов. Собственно, вот уже 20 лет идет настоящий бум ускоренных методов, см., например, [51; 114; 115]. Новые тонкие результаты о сходимости таких методов появляются и по сей день, см., например, [116]. Отметим, что ускоренные методы предложены для задач с ограничениями, с более общим понятием проектирования (неевклидова), для задач со структурой (в условиях модельной общности), в условиях неточности градиента и неточности проектирования, см. [78; 117—119] и цитированную тут литературу. Например, в малоизвестной работе Б.Т. Поляка [14] было показано, что при наличии малого (сколь угодно малого, но фиксированного по масштабу) аддитивного враждебного шума в градиенте [2]

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \delta$$

нельзя гарантировать сходимость градиентного спуска. Более того, можно привести пример, когда он будет расходиться. Однако за счет ранней остановки можно решить данную проблему (см. также работу [101], в которой обсуждается ранняя остановка для метода сопряженных градиентов в условиях неточной информации). В частности, аналогичный результат имеет место и для ускоренных методов. Грубо говоря, можно показать, что по аналогии с (59) имеет место:

$$f(x^N) - f(x_*) \lesssim \frac{LR^2}{N} + R\delta$$

до тех пор (для таких N) пока первое слагаемое не станет меньше второго. В момент, когда это произойдет нужно останавливать метод, иначе он может уже начать расходиться. При этом если концепция шума относительная [2]

$$\|\tilde{\nabla} f(x) - \nabla f(x)\|_2 \leq \alpha \|\nabla f(x)\|_2,$$

где $\alpha \in [0, 1)$, то в отличие от неускоренных методов (см. раздел 3.1) для ускоренных возможность сохранения порядка скорости сходимости остается только при достаточно малых α для сильно выпуклых задач и должным образом убывающих на итерациях α_k для выпуклых задач [119; 120].

Сложно переоценить роль ускорения в современной оптимизации. Практически все современные методы решения задач больших размерностей (в том числе для многих невыпуклых задач) используют ускорение. Ускорение можно дополнительно структурировать. Например, задачу вида

$$\min_x \; f(x) + g(x),$$

в случае когда $f(x)$ – μ_f-сильно выпуклая и L_f-гладкая, а $g(x)$ – μ_g-сильно выпуклая и L_g-гладкая, можно решить с относительной точностью ε за $\mathcal{O}\left(\sqrt{L_f/(\mu_f + \mu_g)}\ln(1/\varepsilon)\right)$ вычислений ∇f и $\mathcal{O}\left(\sqrt{L_g/(\mu_f + \mu_g)}\ln(1/\varepsilon)\right)$ вычислений ∇g [114; 121]. Если использовать стандартное ускорение (не учитывая структуру задачи), то согласно (61) удалось бы получить только такой результат: $\mathcal{O}\left(\sqrt{(L_f + L_g)/(\mu_f + \mu_g)}\ln(1/\varepsilon)\right)$ вычислений ∇f и ∇g, что, очевидно, может быть хуже.

Задачу вида

$$\min_{x,y} f(x, y),$$

в случае когда $f(x, y)$ – выпуклая функция по совокупности аргументов, μ_x-сильно выпуклая и L_x-гладкая по x, а также μ_y-сильно выпуклая и L_y-гладкая по y, можно решить с относительной точностью ε за $\mathcal{O}\left(\sqrt{L_x/\mu_x}\ln(1/\varepsilon)\right)$ вычислений $\nabla_x f$ и $\mathcal{O}\left(\sqrt{L_y/\mu_y}\ln(1/\varepsilon)\right)$ вычислений $\nabla_y f$ [122]. Если использовать стандартное ускорение (не учитывая структуру задачи), то (при некоторых дополнительных оговорках) согласно (61) удалось бы получить только такой результат: $\mathcal{O}\left(\sqrt{\max\{L_x, L_y\}/\min\{\mu_x, \mu_y\}}\ln(1/\varepsilon)\right)$ вычислений $\nabla_x f$ и $\nabla_y f$, что также может быть хуже.

Более того, недавно были получены нетривиальные варианты ускоренных методов для седловых задач со структурой [123]. Например, если рассматривается такая задача

$$\min_x \max_y \; p(x) + F(x, y) - q(y),$$

где $p(x),\ q(y)$ – выпуклые и $L_p,\ L_q$-гладкие функции, а $F(x, y)$ – L_F-гладкая, μ_x-сильно выпуклая и μ_y-сильно вогнутая, то существует такой ускоренный метод, который решит

задачу с относительной точностью ε по зазору двойственности [124] за

$$\mathcal{O}\left(\left(\sqrt{\frac{L_p}{\mu_x}}+\sqrt{\frac{L_q}{\mu_y}}\right)\log\frac{1}{\varepsilon}\right)$$

вычислений $\nabla p(x)$, $\nabla q(y)$ и

$$\mathcal{O}\left(\max\left\{\sqrt{\frac{L_p}{\mu_x}},\sqrt{\frac{L_q}{\mu_y}},\frac{L_F}{\sqrt{\mu_x\mu_y}}\right\}\log\frac{1}{\varepsilon}\right)$$

вычислений $\nabla F(x,y)$.

Начиная с пионерской работы Ю.Е. Нестерова и Б.Т. Поляка 2006 года [7] в нескольких ведущих мировых центрах по оптимизации стали активно разрабатываться тензорные методы (методы, использующие старшие производные). В частности, в работе [125] было показано, что при естественных условиях тензорные методы второго и третьего порядка (использующие производные второго и третьего порядка) могут быть реализованы практически с той же стоимостью итерации, как у метода Ньютона. Оптимальные (с точностью до числовых множителей) ускоренные тензорные методы были предложены в работах [51; 126—130]. Например, оптимальный тензорный метод, использующий производные порядка $r \geq 2$, при условии, что целевая функция μ-сильно выпуклая и тензор r-х производных M_r-Липшицев, для достижения относительной точности ε требует:

$$\mathcal{O}\left(\left(\frac{M_r R^{r-1}}{\mu}\right)^{\frac{2}{3r+1}}+\ln\ln\left(\frac{1}{\varepsilon}\right)\right)$$

итераций (вычислений старших производных). При этом при $r = 2,\ 3$ трудозатратность каждой итерации оптимальных методов практически такая же как и у метода Ньютона. Отметим, что итерационная (оракульная) сложность (необходимое число итераций) во втором слагаемом отвечает локальной итерационной сложности метода Ньютона. Проблема с методом Ньютона только в том, что нужно оказаться в окрестности квадратичной скорости сходимости. Первое слагаемое в приведенной оценке как раз отвечает за время попадания в эту окрестность. Обратим внимание, что даже если доступны старшие производные высокого порядка (r – большое) и целевая функция достаточно гладкая, то второе слагаемое остается по порядку неизменным. Это означает, что локальная скорость сходимости метода Ньютона по порядку оптимальна в классе всевозможных численных методов оптимизации [52].

Из написанного выше может создаться впечатление, что ускорение возможно заточить под структуру любой гладкой задачи. В разделе 3.1 отмечалось, что если вместо (сильной) выпуклости имеет место условие Поляка–Лоясиевича, то в общем случае ускорение невозможно. Кажется, что если ограничиться только (сильно) выпуклыми постановками, то ускорение всегда можно сделать. Как правило, так оно и есть. Но имеются такие постановки выпуклых задач, в которых доказано, что ускорение в общем случае также невозможно. К таким примерам относится задача получения для гладких выпуклых задач оптимальных (по числу коммуникаций и оракульных вызовов — вычислений градиентов) децентрализованных ускоренных методов на меняющихся графах. Децентрализованная оптимизация [131; 132] является частью распределенной оптимизации.

Отметим, что одна из первых работ по распределенной оптимизации была у аспиранта Б.Т. Поляка в конце 70-х годов прошлого века [133]. Бурно развивающийся подраздел распределенной оптимизации – децентрализованная оптимизация на меняющихся со временем графах [134]. Оказывается (см. [135]), что в оценке числа коммуникаций невозможно в общем случае ускорение потому как входит худшее (на итерациях) число обусловленности коммуникационного графа (если бы граф не менялся ускорение было бы возможно за счет ускоренного консенсуса – Чебышёвское ускорение). Несмотря на приведенный пример, все же во всех известных современных постановках выпуклых гладких задач (с различной структурой), как правило, удается добиться ускорения, учитывающего данную структуру. Причем, основные продвижения здесь были получены буквально в последние десять–пятнадцать лет.

Также может показаться, что уже не осталось открытых вопросов, и на все вопросы для задач гладкой выпуклой оптимизации получены ответы. На самом деле, это не так. Например, в упомянутой ранее седловой задаче со структурой открытым остается вопрос о возможности дополнительного разделения оракульных сложностей по числу вызовов $\nabla p(x)$ и $\nabla q(x)$. Но можно не опускаться в такие частности и заметить, что даже для самой обычной задачи гладкой сильно выпуклой оптимизации до сих пор не предложен ускоренный метод, адаптивный по всем параметрам μ, L. При этом в классе неускоренных методов наискорейший спуск решает данную проблему, см. также [136]. К сожалению, адаптивный метод сопряженных градиентов в форме (58), как уже отмечалось, не обязан сходиться с желаемой скоростью на классе (сильно) выпуклых гладких задач. И хотя сейчас есть кандидаты (адаптивные ускоренные методы), которые, возможно, обладают желаемыми свойствами [137], однако пока это не удалось строго доказать.

4. Методы стохастической оптимизации

Задачей стохастической оптимизации называется задача вида

$$\min_{x \in Q} f(x) := \mathbb{E}_\xi f(x, \xi), \tag{65}$$

в которой можно в любой точке x вызвать оракул (подпрограмму), выдающий $\nabla f(x, \xi)$ с новой независимой реализацией ξ (допускается вызов оракула в одной точке x многократно – см. далее батчирование). При этом $\mathbb{E}_\xi \nabla f(x, \xi) \equiv \nabla f(x)$. Целью является найти ε-приближенное решение задачи (65) (по функции $f(x)$) за наименьшее число вызовов оракула. Без преувеличения можно сказать, что современный анализ данных в алгоритмической своей части – это решение соответствующих задач стохастической оптимизации [138].

4.1. Стохастический градиентный спуск

По аналогии с градиентным спуском для решения (65), естественно, рассматривать, так называемые, стохастические градиентные спуски (SGD) [139; 140]:

$$x^{k+1} = \pi_Q \left(x^k - \gamma_k \nabla_x f(x^k, \xi^k) \right), \tag{66}$$

где π_Q – обычное (евклидово) проектирование на множество Q, а ξ^k выбирается независимо от $\xi^0, ..., \xi^{k-1}$.

Если $f(x)$ – выпуклая, то, выбирая $\gamma_k \equiv \frac{R}{M\sqrt{N}}$ можно получить

$$\mathbb{E}f(\bar{x}^N) - f(x^*) \leq \frac{MR}{\sqrt{N}},$$

где $R = \|x^0 - x^*\|_2$ (если x_* не единственное, то в этой формуле можно выбирать ближайшее по 2-норме к x^0), $\mathbb{E}_\xi \|\nabla f(x,\xi)\|_2^2 \leq M^2$ при $x \in Q$ (можно сузить на пересечение Q с некоторым шаром с центром в x^0 и радиусом порядка $2R$ [141] – аналогичное замечание можно делать и относительно всех остальных констант, вводимых далее), $\bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k$. Если $f(x)$ – μ-сильно выпуклая функция, то, выбирая $\gamma_k = \frac{1}{\mu(k+1)}$, можно получить

$$\frac{\mu}{2}\mathbb{E}\|\bar{x}^N - x_*\|_2^2 \leq \mathbb{E}f(\bar{x}^N) - f(x^*) \leq \frac{M^2}{\mu N}. \tag{67}$$

Приведенные оценки в общем случае (без дополнительных предположений) не могут быть улучшены [52]. То есть не существует другого способа агрегирования выборки $\left\{\xi^k\right\}_{k=1}^N$, который давал бы оценки лучше (с точностью до числового множителя) приведенных. Для невыпуклой $f(x)$ гарантировать сходимость к глобальному минимуму уже нельзя. Тем не менее, на практике (66) и его вариации активно применяется и для невыпуклых задач, например, обучения нейронных сетей.

Отметим, что для $Q = \mathbb{R}^n$, видоизменив сам метод (66), при некоторых дополнительных условиях результат (67) можно уточнить следующим образом [142] (приведенная оценка также будет неулучшаемой)

$$\mathbb{E}\|\bar{x}^N - x_*\|_2^2 \lesssim \frac{\mathrm{Tr}\left(\left[\nabla^2 f(x_*)\right]^{-1} \Sigma \left[\nabla^2 f(x_*)\right]^{-1}\right)}{N} + O\left(\frac{1}{N^{3/2}}\right), \tag{68}$$

где $\Sigma = \mathbb{E}_\xi\left[\nabla f(x_*,\xi)\nabla f(x_*,\xi)^T\right]$. Собственно, в этом месте остановимся, чтобы поподробнее рассказать о вкладе Б.Т. Поляка в получении такого рода результатов. В 70-е годы прошлого века Б.Т. Поляк совместно с Я.З. Цыпкиным исследовали следующие псевдоградиентные процедуры стохастического агрегирования (то есть алгоритмы решения задачи (4))

$$x^{k+1} = x^k - \gamma_k \phi(\nabla_x f(x^k, \xi^k)),$$

в которых за счет выбора вектор-функции $\phi(z)$ хотелось получить как можно лучшую скорость сходимости. Базируясь на результатах [143] в работе [13] при аддитивном шуме: $\nabla f(x,\xi) = \nabla f(x) + \xi$ удалось показать, что оптимальным будет такой выбор $\gamma_k = k^{-1}$, $\phi(z) = \left[\nabla^2 f(x_*)\right]^{-1} J^{-1} \nabla \ln p_\xi(z)$, где $p_\xi(z)$ – функция плотности распределения случайного вектора ξ, а информационная матрица Фишера считается по формуле $J = \int \nabla \ln p_\xi(z) \left[\nabla \ln p_\xi(z)\right]^T p_\xi(z) dz$. Под оптимальностью понимается следующее: при $N \to \infty$ имеет место центральная предельная теорема (ЦПТ) в форме:

$$\sqrt{N}\left(x^N - x_*\right) \in \mathcal{N}\left(0, \left[\nabla^2 f(x_*)\right]^{-1} J \left[\nabla^2 f(x_*)\right]^{-1}\right),$$

и при указанном выше способе выбора $\phi(z)$ ковариационная матрица наименьшая (в смысле полуопределенного отношения частичного порядка).

Однако во многих реальных приложениях плотность распределения $p_\xi(z)$ неизвестна. Поэтому использовать ее при выборе $\phi(z)$ нежелательно. Это приводит к корректировке

оптимальной процедуры $\phi(z) = \left[\nabla^2 f(x_*)\right]^{-1} z$ и корректировке основного результата (ЦПТ): при $N \to \infty$

$$\sqrt{N}\left(x^N - x_*\right) \in \mathcal{N}\left(0, \left[\nabla^2 f(x_*)\right]^{-1} \Sigma \left[\nabla^2 f(x_*)\right]^{-1}\right). \tag{69}$$

Здесь использовалось, что $\nabla f(x_*) = 0$. Отметим, что аддитивность шума ξ при этом не требуется. Этот результат также будет оптимальный в классе методов без доступа к $p_\xi(z)$. Однако даже в такой формулировке результат едва ли можно назвать практичным, поскольку для задания $\phi(z)$ требуется знать $\nabla^2 f(x_*)$, что возможно, в основном, только для задач квадратичной стохастической оптимизации. Ключевое наблюдение, позволяющее решить отмеченную проблему, пришло в голову Борису Теодоровичу в конце 80-х годов во сне, и оказалось удивительным по простоте [5; 6]: $\phi(z) = z$, $\gamma_k \sim k^{-\eta}$, $\eta \in (1/2, 1)$, и в качестве выхода алгоритма предлагается использовать не x^N, а $\bar{x}^N$. В этом случае (69) (с заменой x^N на $\bar{x}^N$) останется верным. Таким образом, было показано, как избавиться от типично недоступного предобуславливателя $\left[\nabla^2 f(x_*)\right]^{-1}$. Аналогичное можно проделать и для стохастических вариационных неравенств [144]. Асимптотический вариант (68) очевидным образом получается из (69). Получение неасимптотического варианта требует больших усилий (на появление первых таких результатов ушло еще более 20 лет [145]).

Отметим, что близкая идея (однако реализованная в существенно меньшей общности) использования $\bar{x}^N$ вместо x^N независимо была приблизительно в то же время предложена и на западе Д. Руппертом [146].

Для ряда постановок задач, например, когда множество Q является симплексом, выгоднее (с точки зрения того, как в итоговую оценку будет входить размерность n посредством M и R) использовать неевклидово проектирование (в частности, для симплекса лучше использовать проектирование согласно дивергенции Кульбака–Ляйблера, которое приводит к экспоненциальному взвешиванию компонент стохастического градиента). Соответствующие обобщения SGD принято назвать стохастический метод зеркального спуска (stochastic mirror descent – SMD) [52; 147]. Проблему неадаптивного выбора шага γ_k (требуется заранее знать N) в выпуклом случае решает вариация SMD – стохастический метод двойственных усреднений (stochastic dual averaging method) [148]. Однако более изящно проблема выбора шага решается в AdaGrad версии SGD [149], в которой

$$\gamma_k = \frac{R}{\sqrt{\sum_{j=1}^k \|\nabla f(x^j, \xi^j)\|_2^2}}.$$

При таком выборе шага не требуется и знание глобальной константы M. В современных работах избавляются также и от зависимости R в шаге (класс Parameter-free SGD, к которому относятся, например, DoG [150] и конструкция Mechanic [151]). К сожалению, в общем сильно выпуклом случае пока неизвестно, как можно было бы избавиться от необходимости знания μ (продвижения имеются лишь в частных случаях, например, когда $f(x^*)$ известно). Напомним, что аналогичная проблема была и для ускоренных детерминированных методов, см. конец раздела 3.3.

В случае, если дополнительно известно, что функция $f(x)$ – гладкая (имеет Липшицев градиент), то SGD можно существенно ускорить за счет батч-параллелизации (замены стохастического градиента на выборочное среднее стохастических градиентов

на независимых реализациях):

$$\nabla f(x,\xi) \to \frac{1}{b}\sum_{i=1}^{b} \nabla f(x,\xi^i),$$

где ξ^i – независимые одинаково распределенные, как ξ, а $b \geq 1$ – размер батча, который можно вычислять параллельно. Действительно, рассмотрим, следуя Б.Т. Поляку [2], более точную оценку скорости сходимости SGD в гладком случае (в [2] используется глобальная оценка дисперсии σ^2, однако, несложно показать, что достаточно использовать введенную далее дисперсию в решении σ_*^2, см., например, [152]):

$$\mathbb{E}\left[\|x^N - x_*\|_2^2\right] \leq \|x^0 - x_*\|_2^2 (1-\gamma\mu)^N + \frac{2\gamma\sigma_*^2}{\mu}, \tag{70}$$

где $\sigma_*^2 = \mathbb{E}_\xi\left[\|\nabla f(x_*,\xi) - \nabla f(x_*)\|_2^2\right]$, $\|\nabla f(y,\xi) - \nabla f(x,\xi)\|_2 \leq L\|y-x\|_2$, $\gamma_k \equiv \gamma \leq 1/(2L)$. Несложно также показать, что при выборе $\gamma \equiv 1/(2L)$ за счет батчирования ($\sigma_*^2 \to \sigma_*^2/b$) можно выравнять оба слагаемых в правой части (70), и получить такую версию (60) (тут N – число вычислений $\nabla f(x,\xi)$):

$$\mathbb{E}\|x^N - x_*\|_2^2 \lesssim R^2 \exp\left(-\frac{\mu}{2L}N\right) + \frac{\sigma_*^2}{\mu^2 N}.$$

При этом для ожидаемой невязки по функции можно получить такую оценку:

$$\mathbb{E}f(x^N) - f(x_*) \lesssim LR^2 \exp\left(-\frac{\mu}{2L}N\right) + \frac{\sigma_*^2}{\mu N}. \tag{71}$$

Отметим, что в последнее время в связи с обучением нейронных сетей огромных размеров возникает потребность в изучении роли перепараметризации, что можно сформулировать как малость дисперсии σ_*^2. Современное состояние развития этого направления для неускоренных стохастических градиентных методов описано, например, здесь [153]. Малость σ_*^2 – означает линейную скорость сходимости в небольшую окрестность решения. Такую картину, наверняка, многие наблюдали, на практике, решая задачи обучения. А именно, если выбирать шаг γ (learning rate) достаточно большим, то в ряде случаев можно наблюдать линейную скорость сходимости SGD. Но чем больше шаг γ, тем больше окрестность, внутри которой метод перестает сходиться. Для дальнейшего продвижения требуется уменьшение шага или батчирование.

Многое из того, что написано выше без каких-то существенных изменений переносится и на стохастические вариационные неравенства (седловые задачи), см., например, обзор [124], написание которого было инициировано Б.Т. Поляком летом 2022 года. Насколько нам известно, этот обзор, по-видимому, является последней научной работой Бориса Теодоровича.

4.2. Ускоренные версии стохастического градиентного спуска

Прежде всего заметим, что (71) в форме

$$\mathbb{E}f(x^N) - f(x_*) \lesssim LR^2 \exp\left(-\frac{\mu}{2L}N\right) + \frac{\sigma^2}{\mu N},$$

где

$$\mathbb{E}_\xi \left[\|\nabla f(x,\xi) - \nabla f(x)\|_2^2\right] \leq \sigma^2$$

справедливо при более слабом предположении о Липшицевости градиента

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2.$$

В таких же условиях можно улучшить (ускорить) оценку (71) если за основу брать ускоренный детерминированный метод и заменять в нем градиент на должным образом пробатченный стохастический градиент, см., например, [26; 114; 154], простое изложение имеется в [83; 155] (следует сравнить с (61)):

$$\mathbb{E}f(x^N) - f(x_*) \lesssim LR^2 \exp\left(-\sqrt{\frac{\mu}{4L}}N\right) + \frac{\sigma^2}{\mu N}. \tag{72}$$

Аналогично для выпуклого случая (следует сравнить с (59)):

$$\mathbb{E}f(x^N) - f(x_*) \lesssim \frac{LR^2}{N^2} + \frac{\sigma^2 R^2}{\sqrt{N}}. \tag{73}$$

Если дополнительно известно, что

$$\|\nabla f(y,\xi) - \nabla f(x,\xi)\|_2 \leq L\|y - x\|_2,$$

то приведенные оценки можно уточнить следующим образом [156; 157] (здесь, как и раньше, b – размер батча, только сейчас мы явно его прописываем, поскольку батчированию тут поддаются слагаемые не только содержащие σ_*^2):

$$\mathbb{E}f(x^N) - f(x_*) \lesssim LR^2 \exp\left(-\sqrt{\frac{\mu}{4L}}N\right) + LR^2 \exp\left(-\frac{\mu}{2L}bN\right) + \frac{\sigma_*^2}{\mu bN},$$

$$\mathbb{E}f(x^N) - f(x_*) \lesssim \frac{LR^2}{N^2} + \frac{LR^2}{bN} + \frac{\sigma_*^2 R^2}{\sqrt{bN}}.$$

Причем все приведенные выше оценки в разделе 4.2 имеют место и для задач с ограничениями простой структуры и для неевклидова проектирования. Более того, все эти оценки – оптимальны, то есть не могут быть в общем случае улучшены (с точностью до числовых множителей, в том числе в показатели экспоненты).

Также как и для обычного SGD на практике большую роль играет адаптивность метода. Добавление различных вариантов моментного ускорения к адаптивным методам (типа AdaGrad), упомянутым в разделе 4.1, порождает популярную линейку современных методов типа Adam, AdamW, RMSProp, AdaDelta и т.д., активно использующихся для обучения нейронных сетей. Для выпуклых постановок задач имеется и теоретическое обоснование [158; 159]. Однако вопрос о создании полностью адаптивного ускоренного метода для решения задач выпуклой стохастической оптимизации, насколько нам известно, пока окончательно еще не решен.

В предположении $Q = \mathbb{R}^n$ отметим концепцию мультипликативных помех, развиваемую в работах Б.Т. Поляка в 70-80-е годы прошлого века [2; 3]. На современный манер условие, которому удовлетворяют введенные помехи, можно было бы называть условием сильного роста (strong growth):

$$\mathbb{E}\|\nabla f(x,\xi)\|_2^2 \leq \rho_{sg}\|\nabla f(x)\|_2^2 + \sigma_{sg}^2, \quad \rho_{sg}, \sigma_{sg} \geq 0. \tag{74}$$

Такому условию в гладком случае, например, удовлетворяет координатные методы [160], где рандомизация в стохастическом градиенте возникает за счет случайного выбора координаты, по которой считается частная производная вместо вычисления полного градиента, при этом можно выбирать сразу несколько координат (батч) и сэмплировать не обязательно равномерно, а исходя из свойств производных по направлению [161; 162]. Также под неравенство (74) подходят градиенты, к которым применяется оператор сжатия [163]. Такого рода рандомизация используется в распределенной оптимизации для передачи меньшего числа информации. К примерам операторов сжатия относятся и уже упомянутый выше случайный выбор координат, различные рандомизированные квантизации и округления [164].

Для неускоренных методов, использующих стохастический градиент вида (74), начало построения теории было заложено в уже упомянутых работах [2; 3]. В связи с активным развитием машинного обучения стохастические методы оптимизации стали широко исследоваться в сообществе, в частности, было переоткрыто и предположение сильного роста [165]. На данный момент для неускоренных методов, например, для классического SGD вида (66) имеется хорошо разработанная теория сходимости – см., например, обзорную работу [153]. В частности, для L-гладкой выпуклой целевой функции f справедлива следующая оценка скорости сходимости после N итераций SGD:

$$\mathbb{E}f(\bar{x}^N) - f(x_*) \lesssim \frac{\rho_{sg} L \|x^0 - x_*\|_2^2}{N} + \frac{\sigma_{sg}\|x^0 - x_*\|_2}{\sqrt{N}},$$

где $\bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k$. Если функция является не просто выпуклой, а μ-сильно выпуклой, то можно улучшить оценку и получить, что

$$\mathbb{E}\|x^N - x_*\|_2^2 \lesssim \exp\left(-\frac{\mu N}{\rho_{sg} L}\right)\|x^0 - x_*\|_2^2 + \frac{\sigma_{sg}^2}{\mu^2 N}.$$

Для ускоренных вариантов теория немного беднее, но основные результаты уже были получены [166]. Отметим, что классический ускоренный метод [113] не подходит для такой постановки (74) и необходимо использовать дополнительный моментный член (momentum term) [160; 166]. Тогда в предположении о L-гладкости и выпуклости функции f можно получить следующую оценку скорости сходимости:

$$\mathbb{E}f(\bar{x}^N) - f(x_*) \lesssim \frac{\rho_{sg}^2 L \|x^0 - x_*\|_2^2}{N^2} + \frac{\sigma_{sg}\|x^0 - x_*\|_2}{\sqrt{N}},$$

а для μ-сильно выпуклой функции:

$$\mathbb{E}\|x^N - x_*\|_2^2 \lesssim \exp\left(-\sqrt{\frac{\mu N^2}{4\rho_{sg}^2 L}}\right)\|x^0 - x_*\|_2^2 + \frac{\sigma_{sg}^2}{\mu^2 N}.$$

Отметим, что приведенные результаты удалось с некоторыми оговорками и ослаблением перенести на марковский шум [167].

Между тем, легко заметить, что предположение (74) можно релаксировать до условия слабого роста (weak growth):

$$\mathbb{E}\|\nabla f(x,\xi)\|_2^2 \le \rho_{wg}(f(x) - f(x_*)) + \sigma_{wg}^2 \quad \rho_{wg}, \sigma_{wg} \ge 0. \tag{75}$$

Если выполнено (74), то для выпуклой и L-гладкой функции $\rho_{wg} = 2L\rho_{sg}$ и $\sigma_{wg} = \sigma_{sg}$. Условие (75) является не менее распространенным. В частности, одним из популярных примеров применимости (75) является гладкость в среднем, а именно, нам необходимо предположить, что для любой реализации ξ функция $f(\cdot, \xi)$ является $L(\xi)$-гладкой и выпуклой, и отсюда получить:

$$\begin{aligned}\mathbb{E}\|\nabla f(x,\xi)\|_2^2 \leq & 2\mathbb{E}\|\nabla f(x,\xi) - \nabla f(x_*,\xi)\|_2^2 + 2\mathbb{E}\|\nabla f(x_*,\xi)\|_2^2 \\ \leq & 2\mathcal{L}^2(f(x) - f(x_*)) + 2\mathbb{E}\|\nabla f(x_*,\xi)\|_2^2,\end{aligned} \tag{76}$$

где $\mathcal{L}^2 = \mathbb{E}[L^2(\xi)]$. Отметим, что константа $\mathcal{L}$ может быть значительно хуже, чем L – константа гладкости (Липшицевости градиента) f. Выкладка (76) является самым популярным в литературе примером предположения (75). В частности, оно появляется в работе [168], где авторы прежде всего мотивируются классической задачей наименьших квадратов. В дальнейшем исследование предположений (75) и (76) было обобщено на неравномерную рандомизацию, которая учитывает свойства батчей. В работе [169] предлагается довольно исчерпывающая теория для рандомизации вида (76) с разбором большого числа частных случаев. А именно, классический SGD (66) для выпуклой целевой функции f имеет следующие гарантии сходимости:

$$\mathbb{E}f(\bar{x}^N) - f(x_*) \lesssim \frac{\rho_{wg}\|x^0 - x_*\|_2^2}{N} + \frac{\sigma_{wg}\|x^0 - x_*\|_2}{\sqrt{N}},$$

Если функция является дополнительно μ-сильно выпуклой, то можно получить, что

$$\mathbb{E}\|x^N - x_*\|_2^2 \lesssim \exp\left(-\frac{2\mu N}{\rho_{wg}}\right)\|x^0 - x_*\|_2^2 + \frac{\sigma_{sg}^2}{\mu^2 N}.$$

Говоря о предположениях (74) и (75), важно заметить, что для многих частных случаев σ_{sg} и σ_{wg} равны 0, а это может значительно улучшить гарантии сходимости. Здесь можно отметить популярные и довольно часто встречающиеся примеры перепараметризации ($\nabla f(x_*, \xi) = 0$ для всех ξ) [166] и интерполяции ($f(x,\xi) \geq 0$ и $f(x_*,\xi) = 0$ для всех x и ξ) [170]. Также для упомянутых выше координатных методов справедливо, что $\sigma_{sg} = 0$. Но для самых простых методов с сжатием $\sigma_{sg} \neq 0$. Это мотивировало сообщество создать более продвинутые методы, использующие компрессию [171]. Но стохастический градиент в данных подходах не получается описать с помощью (74) и (75). Можно ввести более сложное предположение [172]:

$$\begin{aligned}\mathbb{E}\left[\|\nabla f(x^k,\xi^k)\|_2^2 \mid x^k\right] &\leq 2\rho_{\sigma_k,1}(f(x^k) - f(x_*)) + \rho_{\sigma_k,2}\sigma_k^2 + \sigma_{\sigma_k,1}^2, \\ \mathbb{E}\left[\|\sigma_{k+1}^2\|_2^2 \mid x^k\right] &\leq (1-p)\sigma_k^2 + 2(f(x^k) - f(x_*)) + \sigma_{\sigma_k,2}^2.\end{aligned}$$

С помощью него можно унифицировано анализировать не только многие современные методы с сжатием, но и популярные алгоритмы, использующие технику редукции дисперсии [173—175], а также продвинутые координатные методы [176]. Важной деталью данного предположения является наличие вспомогательной последовательности $\{\sigma_k^2\}$, которая является уникальной для каждого метода. Эта последовательность обладает важным свойством сходимости, которое и позволяет рассмотреть функцию Ляпунова, состоящую из двух частей: классической вида $\|x^k - x_*\|_2^2$ или $f(x^k) - f(x_*)$ и дополнительной, завязанной на σ_k^2. Например, для μ-сильно выпуклой функции метод SGD (66)

с постоянным шагом $\gamma_k = \gamma$ таким, что

$$\gamma \leq \min\left\{\frac{1}{\mu}; \frac{1}{\rho_{\sigma_k,1} + \frac{2\rho_{\sigma_k,3}\rho_{\sigma_k,2}}{p}}\right\},$$

может гарантировать следующую оценку сходимости [172]:

$$\mathbb{E}V_N \lesssim \exp\left(-\min\left\{\gamma\mu; \frac{p}{2}\right\}N\right)V_0 + \frac{\left(\sigma^2_{\sigma_k,1} + \frac{\rho_{\sigma_k,2}\sigma^2_{\sigma_k,2}}{p}\right)\gamma^2}{\min\{\gamma\mu; p\}},$$

где $V_k = \|x^k - x_*\|_2^2 + \frac{\gamma^2\rho_{\sigma_k,2}\sigma_k^2}{2p}$. Похожие результаты имеются и для выпуклой целевой функции f [177], а также для стохастических вариационных неравенств и седловых задач [178; 179]. Насколько нам известно, на данный момент в условиях слабого роста не известно можно ли добиться ускорения, и если можно, то каким образом?

Наряду с предположениями (74) и (75), можно рассмотреть и похожее условие вида

$$\mathbb{E}\|\nabla f(x,\xi)\|_2^2 \leq \rho_x\|x - x_*\|_2^2 + \sigma_x^2.$$

Касательно него можно выделить следующие работы [135; 167; 180; 181].

Все приведенные выше результаты формулировались в терминах сходимости по математическому ожиданию. Для такой сходимости было достаточно ограниченности второго момента стохастического градиента. В действительности, за счет клиппирования на базе описанных методов можно строить робастные версии, которые гарантированно сходятся с такой же скоростью, но уже в терминах вероятностей больших отклонений, причем имеет место почти субгауссовская концентрация [141; 182; 183]. Заметим, что идея клиппирования (нормализации градиента), как способ борьбы с тяжелыми хвостами, в скалярном случае $n = 1$, по-видимому, впервые появилась в 1973 году в работе Б.Т. Поляка и Я.З. Цыпкина [184] как частный случай того, как можно выбирать функцию $\phi(z) = \min\left\{1, \frac{\lambda}{\|z\|_2}\right\}z$, в псевдоградиентной процедуре из раздела 4.1. Отметим, что результаты Поляка–Цыпкина, кратко описанные в разделе 4.1, недавно были перенесены как раз на постановки задач, в которых аддитивный шум ξ имеет тяжелые хвосты распределения, в том числе не предполагающие наличие конечной дисперсии у шума [185].

В заключение раздела отметим, что недавно ускоренные версии тензорных методов типа Нестерова–Поляка были распространены на достаточно гладкие задачи стохастической оптимизации. В частности, для методов второго порядка полученные результаты оптимальны по числу вызовов стохастических градиентов и числу вызовов стохастических гессианов [186].

4.3. Безградиентные методы

Частным случаем стохастики ξ в $\nabla f(x,\xi)$ может быть рандомизация, которая не «дана извне», а привнесена нами самими. Введение в метод рандомизации может иметь разные причины. Например, ярко об этом написано в статье Ю.Е. Нестерова про покомпонентные методы [160] или в фундаментальной статье А.С. Немировского и др. [147] в части рандомизации умножения матрицы на вектор из единичного симплекса. Но,

пожалуй, самый известный пример рандомизации в стохастической оптимизации – это рандомизация суммы: для целевого функционала вида взвешенной суммы в качестве стохастического градиента используется случайно выбранное слагаемое, см., например, [83]. Однако в этом разделе будут описаны, так называемые, безградиентные (поисковые) методы или методы нулевого порядка, в которых рандомизация – это вынужденная мера, связанная с отсутствием необходимой информации. Такие методы периодически встречались в работах Бориса Теодоровича, см., например, [2; 187], и одна из предложенных им конструкций, которая в последнее время вызывает определенный интерес, будет далее изложена.

Прежде всего, рассмотрим выпуклую задачу оптимизации:

$$\min_{x \in Q} f(x), \tag{77}$$

которая существенно отличается от предыдущих постановок задач, в частности от (65), тем, что оракул может выдать только значение целевой функции $f(x)$ в запрошенной точке x. Такой оракул часто упоминается в литературе как оракул нулевого порядка или безградиентный оракул [188]. Из-за невозможности получить информацию о l-ой производной функции (например, градиент функции f) для решения задачи (77) зачастую прибегают к помощи численных методов нулевого порядка, которые основываются на методах первого порядка, заменяя истинный градиент на различные модели аппроксимации градиента [189]. Например, когда функция $f(x)$ является не просто гладкой, а имеет повышенную гладкость, т.е. функция $f : \mathbb{R}^d \to \mathbb{R}$ имеет непрерывные частные производные до l-го порядка включительно и для всех $x, z \in Q$ удовлетворяет условию Гельдера:

$$\left| f(z) - \sum_{0 \le |n| \le l} \frac{1}{n!} D^n f(x)(z-x)^n \right| \le L_\beta \|z - x\|_2^\beta,$$

где $l < \beta$, $L_\beta > 0$, $n = (n_1, ..., n_d)$ – мультииндекс, $n_i \ge 0$ – целые, $n! = n_1! \cdots n_d!$, $|n| = n_1 + \cdots + n_d$, и $\forall v = (v_1, ..., v_d) \in \mathbb{R}^d$, а также $D^n f(x) v^n = \frac{\partial^{|n|} f(x)}{\partial^{n_1} x_1 \cdots \partial^{n_d} x_d} v_1^{n_1} \cdots v_d^{n_d}$, то при создании безградиентного алгоритма важно подобрать такую аппроксимацию градиента, которая будет использовать преимущества повышенной гладкости функции ($\beta \ge 2$, где β – порядок гладкости функции f). Такую оценку производной по направлению предложили в 1990 году Б.Т. Поляк и А.Б. Цыбаков [4], которая в дальнейшем стала называться «ядерная» аппроксимация и активно использоваться [145; 190—192]:

$$\tilde{\nabla} f(x, \mathbf{e}) = d \frac{f(x + \tau r \mathbf{e}) - f(x - \tau r \mathbf{e})}{2\tau} K(r) \mathbf{e}, \tag{78}$$

где $\tau > 0$, $\mathbf{e}$ – равномерно распределенный на $S_2^d(1) := \{x \in \mathbb{R} : \|x\|_2 = 1\}$, r – равномерно распределенный на отрезке $[-1, 1]$, $\mathbf{e}$ и r независимы, $K : [-1, 1] \to \mathbb{R}$ – фиксированная функция (ядро), которая удовлетворяет следующим условиям:

$$\mathbb{E}[K(u)] = 0, \quad \mathbb{E}[uK(u)] = 1, \quad \mathbb{E}[u^j K(u)] = 0, \quad j = 2, ..., l, \quad \mathbb{E}[|u|^\beta |K(u)|] < \infty.$$

Одним из основных достоинств этой аппроксимации градиента является тот факт, что ядерная аппроксимация (78) требует всего два вычисления значения (реализации) функции на итерации, поскольку информация о повышенной гладкости учитывается в «ядре».

Этот факт существенно улучшает оракульную сложность алгоритма, который использует конечно-разностную схему более высокого порядка в качестве оценки градиента [193], поскольку данная аппроксимация требует большего числа вызовов безградиентного оракула на каждой итерации. К 2020 году появились интересные результаты о скорости сходимости для безградиентного алгоритма [4; 145; 194; 195]: Стохастический метод проекции градиента нулевого порядка, описание которого можно найти в Алгоритме 5.

Алгоритм 5 Стохастический метод проекции градиента нулевого порядка

1: **Requires**: Ядро $K : [-1, 1] \to \mathbb{R}$, размер шага η_k, сглаживающий параметр τ_k.
2: **Initialization**: Сгенерировать скалярный числа $r_1, ..., r_N$ равномерно распределенные на отрезке $[-1, 1]$ и вектора $\mathbf{e}_1, ..., \mathbf{e}_N$ равномерно распределенные на единичной Евклидовой сфере $S_2^d(1)$.
3: **for** $k = 1, \ldots, N$ **do**
4: $\quad f_{\xi_k} := f(x_k + \tau_k r_k \mathbf{e}_k) + \xi_k, \quad f_{\xi'_k} := f(x_k - \tau_k r_k \mathbf{e}_k) + \xi'_k$
5: $\quad \tilde{\nabla} f(x_k, \mathbf{e}_k) := \frac{d}{2\tau_k} \left(f_{\xi_k} - f_{\xi'_k} \right) \mathbf{e}_k K(r_k)$
6: $\quad x_{k+1} := \mathrm{Proj}_Q \left(x_k - \gamma_k \tilde{\nabla} f(x_k, \mathbf{e}_k) \right)$
7: **end for**
8: **return** $\{x_k\}_{k=1}^N$.

Как видно из строчки 4 Алгоритма 5 f_ξ выступает в роли безградиентного оракула, где $\xi \neq \xi'$ – стохастический шум, который характеризует конкретную реализацию (т.е. f_ξ – это значение целевой функции на реализации ξ). Именно поэтому строчку 5 называют аппроксимацией градиента с одноточечной обратной связью. Данная концепция стохастического шума [194—197] формально определяется следующим образом: $\mathbb{E}[\xi^2] \leq \tilde{\Delta}^2$ и $\mathbb{E}[\xi'^2] \leq \tilde{\Delta}^2$, $\tilde{\Delta} \geq 0$, а случайные величины ξ и ξ' не зависят от $\mathbf{e}$ и r. Более того, эта концепция шума не требует предположение о нулевом среднем ξ и ξ', поскольку достаточно того, что $\mathbb{E}[\xi \mathbf{e}] = 0$ и $\mathbb{E}[\xi' \mathbf{e}] = 0$. В Таблице 1 представлены результаты работ [145; 194; 195] через зависимости $N(\varepsilon)$ для различных предположений о выпуклости функции (выпуклая/сильно выпуклая функция), где N – число последовательных итераций, совпадающее (с точность до константы) с общим числом обращений к оракулу нулевого порядка $T = 2N$. Все оценки Таблицы 1 соответствуют случаю, когда $\tilde{\Delta}$ не мало.

Таблица 1: Зависимость числа итераций N от желаемой точности задачи ε, размерности d, константы сильной выпуклости μ и порядка гладкости функции β

	Сильно выпуклый случай	Выпуклый случай
Нижние оценки (2020) [4; 194]	$\Omega\left(\min\left\{\frac{d^{1+\frac{1}{\beta-1}} L_\beta^{\frac{2}{\beta-1}} \tilde{\Delta}^2}{(\mu\varepsilon)^{\frac{\beta}{\beta-1}}}, \frac{d^2 R^2 \tilde{\Delta}^2}{\varepsilon^2}\right\}\right)$	$\Omega\left(\min\left\{\frac{d^{1+\frac{1}{\beta-1}} L_\beta^{\frac{2}{\beta-1}} R^{\frac{2\beta}{\beta-1}} \tilde{\Delta}^2}{\varepsilon^{2+\frac{2}{\beta-1}}}, \frac{d^2 R^2 \tilde{\Delta}^2}{\varepsilon^2}\right\}\right)$
Новицкий и др. (2020) [195]	$\tilde{\mathcal{O}}\left(\frac{d^{2+\frac{1}{\beta-1}} L_\beta^{\frac{2}{\beta-1}} \tilde{\Delta}^2}{(\mu\varepsilon)^{\frac{\beta}{\beta-1}}}\right)$	$\tilde{\mathcal{O}}\left(\frac{d^{2+\frac{1}{\beta-1}} L_\beta^{\frac{2}{\beta-1}} R^{\frac{2\beta}{\beta-1}} \tilde{\Delta}^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$
Akhavan et al. (2020)[194]	$\tilde{\mathcal{O}}\left(\frac{d^{2+\frac{2}{\beta-1}} L_\beta^{\frac{2}{\beta-1}} \tilde{\Delta}^2}{(\mu\varepsilon)^{\frac{\beta}{\beta-1}}}\right)$	$\tilde{\mathcal{O}}\left(\frac{d^{2+\frac{2}{\beta-1}} L_\beta^{\frac{2}{\beta-1}} R^{\frac{2\beta}{\beta-1}} \tilde{\Delta}^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$
Bach et al. (2016) [145]	$\mathcal{O}\left(\frac{d^{2+\frac{2}{\beta-1}} L_\beta^{\frac{2\beta}{\beta-1}} \tilde{\Delta}^{\frac{2(\beta+1)}{\beta-1}}}{(\mu\varepsilon)^{\frac{\beta+1}{\beta-1}}}\right)$	$\mathcal{O}\left(\frac{d^{2+\frac{2}{\beta-1}} \left(L_\beta R \tilde{\Delta}^2\right)^{\frac{2\beta}{\beta-1}}}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$

После некоторой «паузы» в 2023 году авторам работы [192] удалось улучшить верх-

нюю оценку для сильно выпуклого случая за счет более качественного анализа оценки смещения ядерной аппроксимации (78), учитывая, что $\kappa_\beta = \int |u|^\beta |K(u)| du$:

$$\left\| \mathbb{E}\left[\tilde{\nabla} f(x, \mathbf{e})\right] - \nabla f(x) \right\|_2 \leq \kappa_\beta \frac{L}{(l-1)!} \cdot \frac{d}{d+\beta-1} \tau^{\beta-1},$$

а также оценки второго момента ядерной аппроксимации (78) с $\kappa = \int K^2(u) du$:

$$\mathbb{E}\left[\|\tilde{\nabla} f(x, \mathbf{e})\|_2^2\right] \leq 4d\mathbb{E}\left[\|\nabla f(x)\|_2^2\right] + 4d\kappa L^2 \tau^2 + \frac{\kappa d^2 \tilde{\Delta}^2}{\tau^2}.$$

Основное преимущество данных оценок состоит в том, что смещение больше не зависит от размерности d асимптотически. Благодаря этому в работе [192] предоставили следующую верхнюю оценку итерационной сложности (совпадает с оракульной сложностью) для сильно выпуклого случая, размерность в которой не зависит от порядка гладкости:

$$N = \mathcal{O}\left(\frac{d^2 L_\beta^{\frac{2}{\beta-1}} \tilde{\Delta}^2}{(\mu\varepsilon)^{\frac{\beta}{\beta-1}}} \right).$$

Нетрудно заметить, что в этих работах идет «борьба» за оптимальную оракульную сложность $T = 2N$. Однако рассматривая безградиентный алгоритм в последнее время авторы уделяют внимание и другим критериям оптимальности [190], а именно оракульная сложность T, число последовательных итераций N и максимально допустимый уровень враждебного шума, при котором всё ещё удается достичь желаемой точности ε. Один из способов улучшения оценок числа последовательных итераций для безградиентного алгоритма – это взять за базу ускоренный алгоритм первого порядка (например, см. [26; 154]) и применить технику батчирования (где B – размер батча), тем самым достигнув оптимальной оценки в итерационной сложности при $B \geq 4\kappa d$ (см. работу [198]): $N \sim \mathcal{O}\left(\varepsilon^{-1/2}\right)$ и получив («конкурирующие» результаты с Таблицей 1) общее число обращений к оракулу в выпуклом случае для любого размера батча B с $\rho_B = \max\left\{1, \frac{4\kappa d}{B}\right\}$:

$$T = N \cdot B = \max\left\{ \mathcal{O}\left(\sqrt{\frac{\rho_B^2 L R^2}{\varepsilon}} B \right), \mathcal{O}\left(\frac{d^2 L_\beta^{\frac{2}{\beta-1}} \tilde{\Delta}^2}{\varepsilon^{2+\frac{2}{\beta-1}}} \right) \right\}.$$

Кроме того исследование вопроса о максимально допустимом уровне враждебного шума, возвращаемым безградиентным оракулом со значением целевой функции, является не менее важным, поскольку в некоторых приложениях (см. например, [199]) чем больше уровень враждебного шума Δ, тем дешевле вызов безградиентного оракула: безградиентный оракул или оракул нулевого порядка в такой концепции шума принимает следующий вид: $f_\delta(x) = f(x) + \delta(x)$, $|\delta(x)| \leq \Delta$, т.е. оракул возвращает значение целевой функции с некоторым ограниченным шумом. Например, в случае повышенной гладкости функции при выполнении условия Поляка–Лоясиевича (32) в работе [191] рассматриваются различные концепции с враждебным шумом. А также демонстрируется показательный результат преимущества рандомизированного алгоритма и эффективность использования ядерной аппроксимации (78). А в случае, когда функция не является гладкой, но гарантируется M-Липшицевость функции $f(x)$, такой что для всех $x, y \in Q$:

$$|f(y) - f(x)| \leq M\|y - x\|_2,$$

существуют работы [155; 200—202], авторы которых предоставили оценки на максимально допустимый уровень шума Δ в различных настройках задачи, совпадающий с верхними границами, полученными в работе [203] для класса выпуклых M-Липшицевых задач оптимизации.

Обзор современного состояния развития безградиентных методов для (сильно) выпуклых задач в условиях шума, см., например, в [190]. Выше был описан лишь один важный, но все же частный сюжет.

Настоящий доклад представляет пополненную расшифровку записи лекции 12 июля 2023 года А.В. Гасникова на Традиционной школе им. Б.Т. Поляка по оптимизации (https://ssopt.org/).

Список литературы

1. *Поляк Б. Т.* Градиентные методы минимизации функционалов // Журнал вычислительной математики и математической физики. — 1963. — Т. 3, № 4. — С. 643—653.
2. *Поляк Б. Т.* Введение в оптимизацию. — М.: Наука, 1983.
3. *Немировский А. С.*, *Поляк Б. Т.*, *Цыпкин Я. З.* Оптимальные алгоритмы стохастической оптимизации при мультипликативных помехах // Доклады Академии наук. Т. 284. — Российская академия наук. 1985. — С. 564—567.
4. *Поляк Б. Т.*, *Цыбаков А. Б.* Оптимальные порядки точности поисковых алгоритмов стохастической оптимизации // Проблемы передачи информации. — 1990. — Т. 26, № 2. — С. 45—53.
5. *Поляк Б. Т.* Новый метод типа стохастической аппроксимации // Автоматика и телемеханика. — 1990. — № 7. — С. 98—107.
6. *Polyak B. T.*, *Juditsky A. B.* Acceleration of stochastic approximation by averaging // SIAM journal on control and optimization. — 1992. — Т. 30, № 4. — С. 838—855.
7. *Nesterov Y.*, *Polyak B. T.* Cubic regularization of Newton method and its global performance // Mathematical Programming. — 2006. — Т. 108, № 1. — С. 177—205.
8. *Поляк Б. Т.* Градиентные методы решения уравнений и неравенств // Журнал вычислительной математики и математической физики. — 1964. — Т. 4, № 6. — С. 995—1005.
9. *Поляк Б. Т.* О некоторых способах ускорения сходимости итерационных методов // Журнал вычислительной математики и математической физики. — 1964. — Т. 4, № 5. — С. 791—803.
10. *Левитин Е. С.*, *Поляк Б. Т.* Методы минимизации при наличии ограничений // Журнал вычислительной математики и математической физики. — 1966. — Т. 6, № 5. — С. 787—823.
11. *Поляк Б. Т.* Минимизация негладких функционалов // Журнал вычислительной математики и математической физики. — 1969. — Т. 9, № 3. — С. 509—521.
12. *Поляк Б. Т.* Метод сопряженных градиентов в задачах на экстремум // Журнал вычислительной математики и математической физики. — 1969. — Т. 9, № 4. — С. 807—821.

13. *Поляк Б. Т.*, *Цыпкин Я. З.* Оптимальные псевдоградиентные алгоритмы адаптации // Автоматика и телемеханика. — 1980. — № 8. — С. 74—84.

14. *Poljak B.* Iterative algorithms for singular minimization problems // Nonlinear Programming 4. — Elsevier, 1981. — С. 147—166.

15. *Poljak B. T.* Sharp minimum // in book "Generalized Lagrangians and applications". — 1982.

16. *Гасников А. В.* Научный путь Бориса Теодоровича Поляка. Оптимизация // Компьютерные исследования и моделирование. — 2023. — Т. 15, № 2. — С. 235—243.

17. *Fradkov A. L.*, *Granichin O. N.* Boris Teodorovich Polyak // Cybernetics and Physics. — 2023. — Т. 12(1).

18. *Polyak B. T.* Subgradient methods: a survey of Soviet research // Nonsmooth optimization. — 1978. — Т. 3. — С. 5—29.

19. *Shor N. Z.* Minimization methods for non-differentiable functions. Т. 3. — Springer Science & Business Media, 2012.

20. *Шор Н. З.* Методы минимизации недифференцируемых функций и их применения. — Киев: Наук. думка, 1979.

21. *Drori Y.*, *Teboulle M.* An optimal variants of Kelley's cutting-plane method // Math. Programming. — 2016. — Т. 160, № 1. — С. 321—351.

22. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence / N. Loizou [и др.] // International Conference on Artificial Intelligence and Statistics. — PMLR. 2021. — С. 1306—1314.

23. *Wang X.*, *Johansson M.*, *Zhang T.* Generalized Polyak step size for first order optimization with momentum // arXiv preprint arXiv:2305.12939. — 2023.

24. *Hazan E.*, *Kakade S.* Revisiting the Polyak step size // arXiv preprint arXiv:1905.00313. — 2019.

25. *Nesterov Y.* Universal gradient methods for convex optimization problems // Mathematical Programming. — 2015. — Т. 152, № 1. — С. 381—404.

26. *Гасников А. В.*, *Нестеров Ю. Е.* Универсальный метод для задач стохастической композитной оптимизации // ЖВМ и МФ. — 2018. — Т. 58, № 1. — С. 51—68.

27. *Jiang X.*, *Stich S. U.* Adaptive SGD with Polyak stepsize and Line-search: Robust Convergence and Variance Reduction // arXiv preprint arXiv:2308.06058v. — 2023.

28. *Поляк Б. Т.* Один общий метод решения экстремальных задач // Докл. АН СССР. — 1967. — Т. 174, № 1. — С. 33—36.

29. *Huang Y.*, *Lin Q.* Single-Loop Switching Subgradient Methods for Non-Smooth Weakly Convex Optimization with Non-Smooth Convex Constraints // arXiv preprint. — 2023.

30. Mirror descent and convex optimization problems with non-smooth inequality constraints / A. Bayandina [и др.] // Lecture Notes in Mathematics. — 2018. — Т. 2227. — С. 181—213.

31. *Lagae S.* New efficient techniques to solve sparse structured linear systems, with applications to truss topology optimization. — 2017.

32. *Nesterov Y.* Subgradient methods for huge-scale optimization problems // Mathematical Programming. — 2014. — Т. 146, № 1/2. — С. 275—297.

33. Адаптивные алгоритмы зеркального спуска в задачах выпуклого программирования с липшицевыми ограничениями / Ф. С. Стонякин [и др.] // Тр. ИММ УрО РАН. — 2018. — Т. 24, № 2. — С. 266—279.

34. Mirror descent for constrained optimization problems with large subgradient values of functional constraints / F. S. Stonyakin [и др.] // Computer Research and Modeling. — 2020. — Т. 12, № 2. — С. 301—317.

35. Адаптивные субградиентные методы для задач математического программирования с квазивыпуклыми функциями / С. С. Аблаев [и др.] // Труды института математики и механики УрО РАН. — 2023. — Т. 29, № 3. — С. 7—25.

36. *Tiapkin D.*, *Gasnikov A.* Primal-dual stochastic mirror descent for MDPs // International Conference on Artificial Intelligence and Statistics. — PMLR. 2022. — С. 9723—9740.

37. Выпуклая оптимизация: учебное пособие / Е. А. Воронцова [и др.]. — Москва: МФТИ, 2021.

38. A Parameter-free and Projection-free Restarting Level Set Method for Adaptive Constrained Convex Optimization Under the Error Bound Condition / Q. Lin [и др.] // arXiv preprint. — 2022.

39. Subgradient methods for sharp weakly convex functions / D. Davis [и др.] // Journal of Optimization Theory and Applications. — 2018. — Т. 179. — С. 962—982.

40. *Duchi J. C.*, *Ruan F.* Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval // Information and Inference: A Journal of the IMA. — 2019. — Т. 8, № 3. — С. 471—529.

41. *Eldar Y. C.*, *Mendelson S.* Phase retrieval: stability and recovery guarantees // Appl. Comput. Harmon. Anal. — 2014. — Т. 36, № 3. — С. 473—494.

42. *Li X.* Nonconvex Robust Low-rank Matrix Recovery // arXiv preprint. — 2018.

43. *Дудов С. И.*, *Осипцев М. А.* Характеризация решения задач сильно-слабо выпуклого программирования // Матем. сб. — 2021. — Т. 212, № 6. — С. 43—72.

44. Incremental Methods for Weakly Convex Optimization / X. Li [и др.] // OPT2020: 12th Annual Workshop on Optimization for Machine Learning. — 2020.

45. *Davis D.*, *Drusvyatskiy D.*, *Paquette C.* The nonsmooth landscape of phase retrieval // IMA Journal of Numerical Analysis. — 2020. — Т. 40, № 4. — С. 2652—2695.

46. *Davis D.*, *Drusvyatskiy D.*, *Kellie M.* Stochastic model-based minimization under high-order growth // arXiv preprint. — 2018.

47. Субградиентные методы для слабо выпуклых и относительно слабо выпуклых задач с острым минимумом / Ф. С. Стонякин [и др.] // Компьютерные исследования и моделирование. — 2023. — Т. 15, № 2. — С. 393—412.

48. *Li Y.*, *Sun Y.*, *Chi Y.* Low-Rank Positive Semidefinite Matrix Recovery from Corrupted Rank-One Measurements // IEEE Transactions on Signal Processing. — 2017. — Т. 65. — С. 397—408.

49. Robust Principal Component Analysis / E. Candes [и др.] // Journal of the ACM. — 2011.

50. A Theory on the Absence of Spurious Solutions for Nonconvex and Nonsmooth Optimization / C. Josz [и др.] // NeurIPS. — 2018. — С. 2441—2449.

51. Lectures on convex optimization. Т. 137 / Y. Nesterov [и др.]. — Springer, 2018.

52. *Немировский А. С., Юдин Д. Б.* Сложность задач и эффективность методов оптимизации. — Наука, 1979.

53. *Евтушенко Ю. Г.* Методы решения экстремальных задач и их применение в системах оптимизации. — 1982.

54. *Su W., Boyd S., Candes E.* A differential equation for modeling Nesterov's accelerated gradient method: theory and insights // Advances in neural information processing systems. — 2014. — Т. 27.

55. *Wilson A. C., Recht B., Jordan M. I.* A Lyapunov analysis of accelerated methods in optimization // The Journal of Machine Learning Research. — 2021. — Т. 22, № 1. — С. 5040—5073.

56. *Lojasiewicz S.* Une propriété topologique des sous-ensembles analytiques réels // Les équations aux dérivées partielles. — 1963. — Т. 117. — С. 87—89.

57. *Leżański T.* ÜBer das Minimumproblem für Funktionale in Banachschen räumen // Mathematische Annalen. — 1963. — Т. 152, № 4. — С. 271—274.

58. *Karimi H., Nutini J., Schmidt M.* Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition // Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16. — Springer. 2016. — С. 795—811.

59. *Liu C., Zhu L., Belkin M.* Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning // arXiv preprint arXiv:2003.00307. — 2020. — Т. 7.

60. *Fatkhullin I., Polyak B.* Optimizing static linear feedback: Gradient method // SIAM Journal on Control and Optimization. — 2021. — Т. 59, № 5. — С. 3887—3911.

61. *Xiao L.* On the convergence rates of policy gradient methods // Journal of Machine Learning Research. — 2022. — Т. 23, № 282. — С. 1—36.

62. *Ding Y., Zhang J., Lavaei J.* On the global optimum convergence of momentum-based policy gradient // International Conference on Artificial Intelligence and Statistics. — PMLR. 2022. — С. 1910—1934.

63. Global convergence of policy gradient methods for the linear quadratic regulator / M. Fazel [и др.] // International conference on machine learning. — PMLR. 2018. — С. 1467—1476.

64. *Yue P., Fang C., Lin Z.* On the lower bound of minimizing polyak-łojasiewicz functions // The Thirty Sixth Annual Conference on Learning Theory. — PMLR. 2023. — С. 2948—2968.

65. *Yang J., Kiyavash N., He N.* Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems // arXiv preprint arXiv:2002.09621. — 2020.

66. *Garg K., Baranwal M.* Fixed-Time Convergence for a Class of Nonconvex-Nonconcave Min-Max Problems // 2022 Eighth Indian Control Conference (ICC). — IEEE. 2022. — С. 19—24.

67. Solving a class of non-convex min-max games using iterative first order methods / M. Nouiehed [и др.] // Advances in Neural Information Processing Systems. — 2019. — Т. 32.

68. *El Ghaoui L.*, *Lebret H.* Robust solutions to least-squares problems with uncertain data // SIAM Journal on matrix analysis and applications. — 1997. — Т. 18, № 4. — С. 1035—1064.

69. *Муратиди А. Я.*, *Стонякин Ф. С.* Правила остановки градиентного метода для седловых задач с двусторонним условием Поляка-Лоясиевича // arXiv preprint arXiv:2307.09921. — 2023.

70. *Бакушинский А. Б.*, *Поляк Б. Т.* О решении вариационных неравенств // Доклады Академии наук. Т. 219. — Российская академия наук. 1974. — С. 1038—1041.

71. *Stonyakin F.*, *Kuruzov I.*, *Polyak B.* Stopping rules for gradient methods for non-convex problems with additive noise in gradient // Journal of Optimization Theory and Applications. — 2023. — С. 1—21.

72. A theoretical and empirical comparison of gradient approximations in derivative-free optimization / A. S. Berahas [и др.] // Foundations of Computational Mathematics. — 2022. — Т. 22, № 2. — С. 507—560.

73. *Conn A. R.*, *Scheinberg K.*, *Vicente L. N.* Introduction to derivative-free optimization. — SIAM, 2009.

74. *Risteski A.*, *Li Y.* Algorithms and matching lower bounds for approximately-convex optimization // Advances in Neural Information Processing Systems. — 2016. — Т. 29.

75. Convex optimization in hilbert space with applications to inverse problems / A. Gasnikov [и др.] // arXiv preprint arXiv:1703.00267. — 2017.

76. *Kabanikhin S. I.* Inverse and ill-posed problems: theory and applications. — de Gruyter, 2011.

77. *Devolder O.*, *Glineur F.*, *Nesterov Y.* First-order methods of smooth convex optimization with inexact oracle // Mathematical Programming. — 2014. — Т. 146. — С. 37—75.

78. *Devolder O.* Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization : дис. ... канд. / Devolder Olivier. — CORE UCLouvain Louvain-la-Neuve, Belgium, 2013.

79. *d'Aspremont A.* Smooth optimization with approximate gradient // SIAM Journal on Optimization. — 2008. — Т. 19, № 3. — С. 1171—1183.

80. *Vasin A.*, *Gasnikov A.*, *Spokoiny V.* Stopping rules for accelerated gradient methods with additive noise in gradient : тех. отч. / Berlin: Weierstraß-Institut für Angewandte Analysis und Stochastik. — 2021.

81. *Емелин И. В.*, *Красносельский М. А.* Правило останова в итерационных процедурах решения некорректных задач // Автоматика и телемеханика. — 1978. — № 12. — С. 59—63.

82. *Carter R. G.* On the global convergence of trust region algorithms using inexact gradient information // SIAM Journal on Numerical Analysis. — 1991. — Т. 28, № 1. — С. 251—265.

83. *Гасников А. В.* Современные численные методы оптимизации. Метод универсального градиентного спуска. — М.: МЦНМО, 2021.

84. *De Klerk E.*, *Glineur F.*, *Taylor A. B.* On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions // Optimization Letters. — 2017. — Т. 11. — С. 1185—1199.

85. *Puchinin S.*, *Stonyakin F.* Gradient-Type Method for Optimization Problems with Polyak-Lojasiewicz Condition: Relative Inexactness in Gradient and Adaptive Parameters Setting // arXiv preprint arXiv:2307.14101. — 2023.

86. Convex optimization: Algorithms and complexity / S. Bubeck [и др.] // Foundations and Trends® in Machine Learning. — 2015. — Т. 8, № 3/4. — С. 231—357.

87. *Cox B.*, *Juditsky A.*, *Nemirovski A.* Decomposition techniques for bilinear saddle point problems and variational inequalities with affine monotone operators // Journal of Optimization Theory and Applications. — 2017. — Т. 172. — С. 402—435.

88. *Гасников А.*, *Гасникова Е.* Модели равновесного распределения транспортных потоков в больших сетях. — 2023.

89. Efficient numerical methods to solve sparse linear equations with application to pagerank / A. Anikin [и др.] // Optimization Methods and Software. — 2022. — Т. 37, № 3. — С. 907—935.

90. *Bomze I. M.*, *Rinaldi F.*, *Zeffiro D.* Frank–Wolfe and friends: a journey into projection-free first-order optimization methods // 4OR. — 2021. — Т. 19. — С. 313—345.

91. Conditional gradient methods / G. Braun [и др.] // arXiv preprint arXiv:2211.14103. — 2022.

92. Zero-Order Stochastic Conditional Gradient Sliding Method for Non-smooth Convex Optimization / A. Lobanov [и др.] // arXiv preprint arXiv:2303.02778. — 2023.

93. *Vedernikov R.*, *Rogozin A.*, *Gasnikov A.* Decentralized conditional gradient method over time-varying graphs // arXiv preprint arXiv:2307.10978. — 2023.

94. Adaptive Variant of the Frank-Wolfe Algorithm for Convex Optimization Problems / G. Aivazian [и др.] // arXiv preprint arXiv:2307.16059. — 2023.

95. *Vial J.-P.* Strong convexity of sets and functions // Journal of Mathematical Economics. — 1982. — Т. 9, № 1/2. — С. 187—205.

96. *Vial J.-P.* Strong and weak convexity of sets and functions // Mathematics of Operations Research. — 1983. — Т. 8, № 2. — С. 231—259.

97. *Половинкин Е. С.* Сильно выпуклый анализ // Математический сборник. — 1996. — Т. 187, № 2. — С. 103—130.

98. *Ito M.*, *Lu Z.*, *He C.* A Parameter-Free Conditional Gradient Method for Composite Minimization under Hölder Condition // Journal of Machine Learning Research. — 2023. — Т. 24. — С. 1—34.

99. *Taylor A. B.*, *Hendrickx J. M.*, *Glineur F.* Smooth strongly convex interpolation and exact worst-case performance of first-order methods // Mathematical Programming. — 2017. — Т. 161. — С. 307—345.

100. Super-acceleration with cyclical step-sizes / B. Goujaud [и др.] // International Conference on Artificial Intelligence and Statistics. — PMLR. 2022. — С. 3028—3065.

101. *Немировский А. С.* О регуляризующих свойствах метода сопряжённых градиентов на некорректных задачах // Журнал вычислительной математики и математической физики. — 1986. — Т. 26, № 3. — С. 332—347.

102. Acceleration methods / A. d'Aspremont, D. Scieur, A. Taylor [и др.] // Foundations and Trends® in Optimization. — 2021. — Т. 5, № 1/2. — С. 1—245.

103. *Scieur D.*, *Pedregosa F.* Universal average-case optimality of polyak momentum // International conference on machine learning. — PMLR. 2020. — С. 8565—8572.

104. *Гельфанд И. М.*, *Цетлин М. Л.* Принцип нелокального поиска в системах автоматической оптимизации // Доклады Академии наук. Т. 137. — Российская академия наук. 1961. — С. 295—298.

105. *Lessard L.*, *Recht B.*, *Packard A.* Analysis and design of optimization algorithms via integral quadratic constraints // SIAM Journal on Optimization. — 2016. — Т. 26, № 1. — С. 57—95.

106. *Ghadimi E.*, *Feyzmahdavian H. R.*, *Johansson M.* Global convergence of the heavy-ball method for convex optimization // 2015 European control conference (ECC). — IEEE. 2015. — С. 310—315.

107. *Goujaud B.*, *Taylor A.*, *Dieuleveut A.* Provable non-accelerations of the heavy-ball method // arXiv preprint arXiv:2307.11291. — 2023.

108. *O'Donoghue B.*, *Candes E.* Adaptive restart for accelerated gradient schemes // Foundations of computational mathematics. — 2015. — Т. 15. — С. 715—732.

109. *Danilova M.*, *Kulakova A.*, *Polyak B.* Non-monotone behavior of the heavy ball method // Difference Equations and Discrete Dynamical Systems with Applications: 24th ICDEA, Dresden, Germany, May 21–25, 2018 24. — Springer. 2020. — С. 213—230.

110. *Немировский А.* Орт-метод гладкой выпуклой минимизации // Изв. АН СССР. Техн. кибернетика. — 1982. — № 2. — С. 18—29.

111. Is local SGD better than minibatch SGD? / B. Woodworth [и др.] // International Conference on Machine Learning. — PMLR. 2020. — С. 10334—10343.

112. The min-max complexity of distributed stochastic convex optimization with intermittent communication / B. E. Woodworth [и др.] // Conference on Learning Theory. — PMLR. 2021. — С. 4386—4437.

113. *Нестеров Ю. Е.* Метод решения задачи выпуклого программирования со скоростью сходимости $O(1/k^2)$ // Доклады Академии наук. Т. 269. — Российская академия наук. 1983. — С. 543—547.

114. *Lan G.* First-order and stochastic optimization methods for machine learning. Т. 1. — Springer, 2020.

115. *Lin Z.*, *Li H.*, *Fang C.* Accelerated optimization for machine learning // Nature Singapore: Springer. — 2020.

116. *Peng W.*, *Wang T.* The Nesterov-Spokoiny Acceleration: $o(1/k^2)$ Convergence without Proximal Operations // arXiv preprint arXiv:2308.14314. — 2023.

117. Inexact model: A framework for optimization and variational inequalities / F. Stonyakin [и др.] // Optimization Methods and Software. — 2021. — Т. 36, № 6. — С. 1155—1201.

118. *Zhang Z.*, *Lan G.* Solving Convex Smooth Function Constrained Optimization Is As Almost Easy As Unconstrained Optimization // arXiv preprint arXiv:2210.05807. — 2022.

119. Accelerated gradient methods with absolute and relative noise in the gradient / A. Vasin [и др.] // Optimization Methods and Software. — 2023. — С. 1—50.

120. Intermediate Gradient Methods with Relative Inexactness / N. Kornilov [и др.] // arXiv preprint arXiv:2310.00506. — 2023.

121. Optimal gradient sliding and its application to optimal distributed optimization under similarity / D. Kovalev [и др.] // Advances in Neural Information Processing Systems. — 2022. — Т. 35. — С. 33494—33507.

122. *Kovalev D.*, *Gasnikov A.*, *Malinovsky G.* An Optimal Algorithm for Strongly Convex Min-min Optimization // arXiv preprint arXiv:2212.14439. — 2022.

123. Optimal Algorithm with Complexity Separation for Strongly Convex-Strongly Concave Composite Saddle Point Problems / E. Borodich [и др.] // arXiv preprint arXiv:2307.12946. — 2023.

124. Smooth monotone stochastic variational inequalities and saddle point problems: A survey / A. Beznosikov [и др.] // European Mathematical Society Magazine. — 2023. — № 127. — С. 15—28.

125. *Nesterov Y.* Implementable tensor methods in unconstrained convex optimization // Mathematical Programming. — 2021. — Т. 186. — С. 157—183.

126. *Monteiro R. D.*, *Svaiter B. F.* An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods // SIAM Journal on Optimization. — 2013. — Т. 23, № 2. — С. 1092—1125.

127. Near optimal methods for minimizing convex functions with lipschitz p-th derivatives / A. Gasnikov [и др.] // Conference on Learning Theory. — PMLR. 2019. — С. 1392—1393.

128. *Kovalev D.*, *Gasnikov A.* The first optimal acceleration of high-order methods in smooth convex optimization // Advances in Neural Information Processing Systems. — 2022. — Т. 35. — С. 35339—35351.

129. Optimal and Adaptive Monteiro-Svaiter Acceleration / Y. Carmon [и др.] // Advances in Neural Information Processing Systems. Т. 35 / под ред. S. Koyejo [и др.]. — Curran Associates, Inc., 2022. — С. 20338—20350. — URL: https://proceedings.neurips.cc/paper_files/paper/202

130. Exploiting higher-order derivatives in convex optimization methods / D. Kamzolov [и др.] // arXiv preprint arXiv:2208.13190. — 2022.

131. *Bertsekas D.*, *Tsitsiklis J.* Parallel and distributed computation: numerical methods. — Athena Scientific, 2015.

132. Recent theoretical advances in decentralized distributed convex optimization / E. Gorbunov [и др.] // High-Dimensional Optimization and Probability: With a View Towards Data Science. — Springer, 2022. — С. 253—325.

133. *Кибардин В. М.* Декомпозиция по функциям в задаче минимизации // Автоматика и телемеханика. — 1979. — № 9. — С. 66—79.

134. Decentralized optimization over time-varying graphs: a survey / A. Rogozin [и др.] // arXiv preprint arXiv:2210.09719. — 2022.

135. Decentralized Optimization Over Slowly Time-Varying Graphs: Algorithms and Lower Bounds / D. Metelev [и др.] // arXiv preprint arXiv:2307.12562. — 2023.

136. *Bao C.*, *Chen L.*, *Li J.* The Global R-linear Convergence of Nesterov's Accelerated Gradient Method with Unknown Strongly Convex Parameter // arXiv preprint arXiv:2308.14080. — 2023.

137. *Guminov S.*, *Gasnikov A.*, *Kuruzov I.* Accelerated methods for weakly-quasi-convex optimization problems // Computational Management Science. — 2023. — Т. 20, № 1. — С. 1—19.

138. Algorithmic stochastic convex optimization / A. Beznosikov [и др.]. — Springer, 2024.

139. *Robbins H.*, *Monro S.* A stochastic approximation method // The annals of mathematical statistics. — 1951. — С. 400—407.

140. *Ермольев Ю.* Методы стохастического программирования. — М.: Наука, 1976.

141. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance / A. Sadiev [и др.] // ICML. — 2023.

142. Root-sgd: Sharp nonasymptotics and asymptotic efficiency in a single algorithm / C. J. Li [и др.] // Conference on Learning Theory. — PMLR. 2022. — С. 909—981.

143. *Невельсон М.*, *Хасьминский Р.* Стохастическая аппроксимация и рекуррентное оценивание. — М.: Наука, 1972.

144. *Fort G.* Central limit theorems for stochastic approximation with controlled Markov chain dynamics // ESAIM: Probability and Statistics. — 2015. — Т. 19. — С. 60—80.

145. *Bach F.*, *Perchet V.* Highly-smooth zero-th order online optimization // Conference on Learning Theory. — PMLR. 2016. — С. 257—283.

146. *Ruppert D.* Efficient estimations from a slowly convergent Robbins-Monro process : тех. отч. / Cornell University Operations Research ; Industrial Engineering. — 1988.

147. Robust stochastic approximation approach to stochastic programming / A. Nemirovski [и др.] // SIAM Journal on optimization. — 2009. — Т. 19, № 4. — С. 1574—1609.

148. *Nesterov Y.* Primal-dual subgradient methods for convex problems // Mathematical programming. — 2009. — Т. 120, № 1. — С. 221—259.

149. *Duchi J.*, *Hazan E.*, *Singer Y.* Adaptive subgradient methods for online learning and stochastic optimization. // Journal of machine learning research. — 2011. — Т. 12, № 7.

150. *Ivgi M.*, *Hinder O.*, *Carmon Y.* DoG is SGD's Best Friend: A Parameter-Free Dynamic Step Size Schedule. — 2023. — arXiv: 2302.12022 [cs.LG].

151. *Cutkosky A.*, *Defazio A.*, *Mehta H.* Mechanic: A Learning Rate Tuner // arXiv preprint arXiv:2306.00144. — 2023.

152. *Stich S. U.* Unified optimal analysis of the (stochastic) gradient method // arXiv preprint arXiv:1907.04232. — 2019.

153. *Gorbunov E.* Unified analysis of SGD-type methods // arXiv preprint arXiv:2303.16502. — 2023.

154. *Lan G.* An optimal method for stochastic composite optimization // Mathematical Programming. — 2012. — Т. 133, № 1/2. — С. 365—397.

155. The power of first-order smooth optimization for black-box non-smooth problems / A. Gasnikov [и др.] // International Conference on Machine Learning. — PMLR. 2022. — С. 7241—7265.

156. *Woodworth B. E.*, *Srebro N.* An even more optimal stochastic optimization algorithm: minibatching and interpolation learning // Advances in Neural Information Processing Systems. — 2021. — Т. 34. — С. 7333—7345.

157. Accelerated stochastic approximation with state-dependent noise / S. Ilandarideva [и др.] // arXiv preprint arXiv:2307.01497. — 2023.

158. Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization / A. Kavis [и др.] // Advances in neural information processing systems. — 2019. — Т. 32.

159. *Ene A.*, *Nguyen H. L.*, *Vladu A.* Adaptive gradient methods for constrained convex optimization and variational inequalities // Proceedings of the AAAI Conference on Artificial Intelligence. Т. 35. — 2021. — С. 7314—7321.

160. *Nesterov Y.* Efficiency of coordinate descent methods on huge-scale optimization problems // SIAM Journal on Optimization. — 2012. — Т. 22, № 2. — С. 341—362.

161. *Richtárik P.*, *Takáč M.* On optimal probabilities in stochastic coordinate descent methods // Optimization Letters. — 2016. — Т. 10. — С. 1233—1243.

162. *Qu Z.*, *Richtárik P.* Coordinate descent with arbitrary sampling I: Algorithms and complexity // Optimization Methods and Software. — 2016. — Т. 31, № 5. — С. 829—857.

163. QSGD: Communication-efficient SGD via gradient quantization and encoding / D. Alistarh [и др.] // Advances in neural information processing systems. — 2017. — Т. 30.

164. On biased compression for distributed learning / A. Beznosikov [и др.] // arXiv preprint arXiv:2002.12410. — 2020.

165. *Schmidt M.*, *Roux N. L.* Fast convergence of stochastic gradient descent under a strong growth condition // arXiv preprint arXiv:1308.6370. — 2013.

166. *Vaswani S.*, *Bach F.*, *Schmidt M.* Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron // The 22nd international conference on artificial intelligence and statistics. — PMLR. 2019. — С. 1195—1204.

167. First Order Methods with Markovian Noise: from Acceleration to Variational Inequalities / A. Beznosikov [и др.] // arXiv preprint arXiv:2305.15938. — 2023.

168. *Moulines E.*, *Bach F.* Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning // Advances in Neural Information Processing Systems. Т. 24 / под ред. J. Shawe-Taylor [и др.]. — Curran Associates, Inc., 2011. — URL: https://proceedings.neurips.cc/

169. SGD: General analysis and improved rates / R. M. Gower [и др.] // International conference on machine learning. — PMLR. 2019. — С. 5200—5209.

170. *Ma S.*, *Bassily R.*, *Belkin M.* The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning // International Conference on Machine Learning. — PMLR. 2018. — С. 3325—3334.

171. Distributed learning with compressed gradient differences / K. Mishchenko [и др.] // arXiv preprint arXiv:1901.09269. — 2019.

172. *Gorbunov E.*, *Hanzely F.*, *Richtárik P.* A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent // International Conference on Artificial Intelligence and Statistics. — PMLR. 2020. — С. 680—690.

173. *Defazio A.*, *Bach F.*, *Lacoste-Julien S.* SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives // Advances in neural information processing systems. — 2014. — Т. 27.

174. *Johnson R.*, *Zhang T.* Accelerating stochastic gradient descent using predictive variance reduction // Advances in neural information processing systems. — 2013. — Т. 26.

175. *Kovalev D.*, *Horváth S.*, *Richtárik P.* Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop // Algorithmic Learning Theory. — PMLR. 2020. — С. 451—467.

176. *Hanzely F.*, *Mishchenko K.*, *Richtárik P.* SEGA: Variance reduction via gradient sketching // Advances in Neural Information Processing Systems. — 2018. — Т. 31.

177. Unified analysis of stochastic gradient methods for composite convex and smooth optimization / A. Khaled [и др.] // Journal of Optimization Theory and Applications. — 2023. — С. 1—42.

178. Stochastic gradient descent-ascent: Unified theory and new efficient methods / A. Beznosikov [и др.] // International Conference on Artificial Intelligence and Statistics. — PMLR. 2023. — С. 172—235.

179. *Maslovskii A. Y.* A unified analysis of variational inequality methods: variance reduction, sampling, quantization, and coordinate descent // Computational Mathematics and Mathematical Physics. — 2023. — Т. 63, № 2. — С. 147—174.

180. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling / Y.-G. Hsieh [и др.] // Advances in Neural Information Processing Systems. — 2020. — Т. 33. — С. 16223—16234.

181. Stochastic extragradient: General analysis and improved rates / E. Gorbunov [и др.] // International Conference on Artificial Intelligence and Statistics. — PMLR. 2022. — С. 7865—7901.

182. Алгоритмы робастной стохастической оптимизации на основе метода зеркального спуска / А. В. Назин [и др.] // Автоматика и телемеханика. — 2019. — № 9. — С. 64—90.

183. High-Probability Convergence for Composite and Distributed Stochastic Minimization and Variational Inequalities with Heavy-Tailed Noise / E. Gorbunov [и др.]. — 2023. — arXiv: 2310.01860 [math.OC].

184. *Поляк Б. Т.*, *Цыпкин Я. З.* Псевдоградиентные алгоритмы адаптации и обучения. Автоматика и телемеханика // Автоматика и телемеханика. — 1973. — № 3. — С. 45—68.

185. Nonlinear gradient mappings and stochastic optimization: A general framework with applications to heavy-tail noise / D. Jakovetić [и др.] // SIAM Journal on Optimization. — 2023. — Т. 33, № 2. — С. 394—423.

186. Advancing the lower bounds: An accelerated, stochastic, second-order method with optimal adaptation to inexactness / A. Agafonov [и др.]. — 2023. — arXiv: 2309.01570 [math.OC].

187. *Граничин О. Н., Поляк Б. Т.* Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах. — М.: Наука, 2003.

188. *Rosenbrock H.* An automatic method for finding the greatest or least value of a function // The computer journal. — 1960. — Т. 3, № 3. — С. 175—184.

189. *Kiefer J., Wolfowitz J.* Stochastic estimation of the maximum of a regression function // The Annals of Mathematical Statistics. — 1952. — С. 462—466.

190. Randomized gradient-free methods in convex optimization / A. Gasnikov [и др.] // arXiv preprint arXiv:2211.13566. — 2022.

191. *Lobanov A., Gasnikov A., Stonyakin F.* Highly Smoothness Zero-Order Methods for Solving Optimization Problems under PL Condition // Журнал вычислительной математики и математической физики. — 2023.

192. Gradient-free optimization of highly smooth functions: improved analysis and a new algorithm / A. Akhavan [и др.] // arXiv preprint arXiv:2306.02159. — 2023.

193. A theoretical and empirical comparison of gradient approximations in derivative-free optimization / A. S. Berahas [и др.] // Foundations of Computational Mathematics. — 2022. — Т. 22, № 2. — С. 507—560.

194. *Akhavan A., Pontil M., Tsybakov A.* Exploiting higher order smoothness in derivative-free optimization and continuous bandits // Advances in Neural Information Processing Systems. — 2020. — Т. 33. — С. 9017—9027.

195. *Novitskii V., Gasnikov A.* Improved exploiting higher order smoothness in derivative-free optimization and continuous bandit // arXiv preprint arXiv:2101.03821. — 2021.

196. *Гасников А., Двуреченский П., Нестеров Ю.* Стохастические градиентные методы с неточным оракулом // Труды Московского физико-технического института. — 2016. — Т. 8, 1 (29). — С. 41—91.

197. *Граничин О. Н., Иванский Ю. В., Копылова К. Д.* Метод Б.Т.Поляка на основе стохастической функции Ляпунова для обоснования состоятельности оценок поискового алгоритма стохастической аппроксимации при неизвестных, но ограниченных помехах // Журнал вычислительной математики и математической физики. — 2023.

198. *Lobanov A., Bashirov N., Gasnikov A.* The Black-Box Optimization Problem: Zero-Order Accelerated Stochastic Method via Kernel Approximation. — 2023. — arXiv: 2310.02371 [math.OC].

199. Learning supervised pagerank with gradient-based and gradient-free optimization methods / L. Bogolubsky [и др.] // Advances in neural information processing systems. — 2016. — Т. 29.

200. Noisy zeroth-order optimization for non-smooth saddle point problems / D. Dvinskikh [и др.] // International Conference on Mathematical Optimization Theory and Operations Research. — Springer. 2022. — С. 18—33.

201. Gradient-Free Federated Learning Methods with l_1 and l_2-Randomization for Non-Smooth Convex Stochastic Optimization Problems / A. Lobanov [и др.] // arXiv preprint arXiv:2211.10783. — 2022.

202. Accelerated Zeroth-order Method for Non-Smooth Stochastic Convex Optimization Problem with Infinite Variance / N. Kornilov [и др.] // arXiv preprint arXiv:2310.18763. — 2023.

203. *Risteski A., Li Y.* Algorithms and matching lower bounds for approximately-convex optimization // Advances in Neural Information Processing Systems. — 2016. — Т. 29.