ФГАОУВО «Московский физико-технический институт (национальный исследовательский университет)»
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»

Работа допущена к защите
зав. кафедрой

_____ Рудаков К. В.

«____» _____ 2020 г.

Выпускная квалификационная работа на соискание степени

БАКАЛАВРА

Тема: **Разработка распределенных децентрализованных безградиентных методов решения негладких задач стохастической выпуклой оптимизации**

Направление: 03.03.01 – Прикладные математика и физика

Выполнил студент гр. 674а _____ Безносиков Александр Николаевич

Научный руководитель,
д. ф.-м. н. _____ Гасников А. В.

Москва – 2020

Аннотация

В этой работе представляется новый метод, основанный на Sliding Algorithm [1, 2], для задачи выпуклой композитной оптимизации, которая состоит из два частей: гладкой и негладкой. Новый метод использует стохастический оракул нулевого порядка для негладкой части и оракула первого порядка для гладкой части. Насколько известно, это первый метод, который использует такой смешанный оракул для задачи композитной оптимизации. В работе доказываются оценки на скорость сходимости нового метода, которые соответствуют скорости для метода первого порядка с точностью до коэффициента, пропорционального размерности пространства или, в некоторых случаях, квадрата логарифма от размерности пространства. Метод применяется для задачи децентрализованной распределенной оптимизации, а также находятся верхние оценки для числа коммуникационных раундов, которые соответствуют лучшим известным оценкам. Более того, полученная оценка числа вызовов оракула нулевого порядка соответствует аналогичной лучшей оценке для децентрализованной распределенной оптимизации первого порядка с коэффициентом, пропорциональным размерности пространства или, в некоторых случаях, даже квадрату логарифма от нее. Метод использует стохастическую аппроксимацию градиента в виде конечных разностей, в этом случае функция должна быть определена не только на множестве оптимизации, но и в некоторой его окрестности. Во второй части работы анализируется случай, когда такое предположение не может быть сделано, предлагается общий подход, как модернизировать метод для решения этой проблемы, а также применяется этот подход к частным случаям некоторых классических множеств.

Ключевые слова: *gradient sliding, оптимизация нулевого порядка, безградиентные оракулы, децентрализованная распределенная оптимизация, композитная оптимизация.*

Оглавление

1. Введение

Актуальность работы. В этой работе рассматривается задачу минимизации конечной суммы функций

$$\min_{x \in X \subseteq \mathbb{R}^n} f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x), \tag{1}$$

где каждая f_i является выпуклой и дифференцируемой, X – замкнутое и выпуклое множество. Задачи такого типа возникают в машинном обучении [3], статистике [4] и теории управления [5]. В частности, интересен случай, когда функции f_i хранятся на различных устройствах (процессорах), которые соединены в сеть [6–13]. Такой сценарий часто используется для ускорения решения задача машинного обучения большого размера, когда целью является ускорение обучения больших моделей или когда информация, определяющая f_i известна только i-ому *устройству* (федеративное обучение).

Задачи распределенной оптимизации делятся на два основных типа: централизованные и децентрализованные.

Централизованную задачу можно описать следующим образом

1) все устройства параллельно производят вычисления градиента или стохастического градиента функции f_i;

2) далее каждое устройство посылает посчитанный градиент на главное устройство – *сервер*;

3) сервер обрабатывает полученную информацию (делает шаг) и передает новую информацию каждому устройству, которая необходима для выполнения новой итерации, а затем процесс повторяется.

Однако такой подход имеет несколько недостатков, например, можно выделить проблему синхронизации или высокие требования к серверу. Есть много работ, которые пытаются решить эти недостатки (смотри, например, [14–17]).

Другой возможный подход для устранения этих недостатков заключается в использовании децентрализованной сети [18]. В данном подходе устройства общаются не с сервером, со своими соседями по сети, и общение происходит одновременно. Отметим, что такой подход является более надежным, например, он может применяться к изменяющимся во времени (беспроводным) сетям связи [19].

Цели работы. В данной работе ставятся следующие цели:

- для задачи композитной оптимизации разработать новый метод, который использует смешанный оракул: для одной из частей задачи – оракул нулевого порядка, а для другой – первого порядка;
- применить полученный метод для задачи децентрализованной распределенной оптимизации.

Базовые методы. За основу используется Sliding Algorithm, предложенный в [1, 2]. В отличие от оригинальных статей наш метод работает с оракулом первого и нулевого порядка. Концепция оракулов нулевого порядка, которая рассматривается в работе, изложена в [20, 21].

Основные положения, выносимые на защиту.

1. Метод для решения выпуклой задачи композитной оптимизации, содержащей негладкую часть и L-гладкую часть, который использует смещенный стохастический оракул нулевого порядка для негладкой компоненты и оракул первого порядка для гладкой компоненты;
2. Модификация метода для сильно выпуклого случая;
3. Техника использования метода для задачи децентрализованной распределенной оптимизации;
4. Техника для работы с оракулами нулевого порядка, когда функция задана только в пределах допустимого множества оптимизации, а также оценки на уровень шума в данном случае.

Научная новизна. В этой работе приводится новый метод под названием zeroth-order Sliding Algorithm (`zoSA`) для задачи композитной оптимизации. Этот метод является первым, который использует оракулы нулевого и первого порядка для композитной задачи.

Теоретическая значимость. Сходимость предложенного метода соответствует известным результатам для числа вызовов оракула, отвечающего за гладкую компоненту. Для негладкой компоненты доказывается, что требуемое количество вызовов оракула нулевого порядка обычно в n раз, а в некоторых случаях в $\log^2 n$

больше, чем соответствующая оценка, полученная для числа вызовы оракула первого порядка, что естественно для оптимизации, не использующей градиенты (смотри, например, [22]). Кроме того, предложенный метод рассматривается в случае, когда гладкий член дополнительно является сильно выпуклым.

Далее `zoSA` применяется к задаче децентрализованной распределенной оптимизации. Полученные результаты соответствуют лучшим результатам для негладкой децентрализованной задачи с точки зрения коммуникационных раундов.

Оракул нулевого порядка, который рассматривается в этой работе, вместо градиента возвращает его аппроксимацию через конечные разности. Концепция использования такого оракула не нова (см. [21, 23]). Для такого оракула необходимо, чтобы функция была определена в некоторой окрестности множества оптимизации, поскольку, когда вычисляется конечная разность, делается небольшой шаг от точки, и этот шаг может вывести нас за пределы множества. Насколько известно, во всех предыдущих работах авторы исходят из того, что упомянутое предположение выполнено. В этой же работе поднимается вопрос о том, что можно сделать, когда функция определена только на заданном множестве из-за некоторых свойств задачи. Приводится общий подход о том, как действовать в том случае, когда нам запрещено выходить за пределы исходного множества оптимизации.

Практическая значимость. Предложенный метод имеет большое практическое значение. В работе приводятся численные эксперименты для некоторых классических задач, в том числе и задач машинного обучения. Результаты показывают, что `zoSA`, который использует смешанный оракул, имеет сходимость лучше, чем зеркальный спуск с оракулом нулевого порядка, а иногда и лучше, чем и зеркальный спуск с оракулом первого порядка.

Степень достоверности и апробация работы. Достоверность результатов подтверждена математическими доказательствами и экспериментальной проверкой полученных методов. Результаты работы были приняты к публикации в изданиях, индексируемых Scorpus:

- IFAC-PapersOnLine,
- Communications in Computer and Information Science (CCIS),

докладывались и обсуждались на следующих научных конференциях:

- 62 научная конференция МФТИ с международным участием, 2019 г,
- Quasilinear Equations, Inverse Problems and Their Applications 2019,

будут представлены на конференциях:

- 21st IFAC World Congress 2020,
- Mathematical Optimization Theory and Operations Research, 2020.

2. Алгоритмы и основные результаты

В этом разделе представлены модификации Sliding Algorithm [6], использующие смешанный оракул, для выпуклых и сильно выпуклых задач, а также приведены теоремы об оценках сложности этих алгоритмов. Доказательства можно найти[1] в [24, 25]. Для удобства все доказательства также приведены в разделах 7-10.

2.1. Обозначения и вспомогательные факты

В работе используется $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$, чтобы определить стандартное скалярное произведение векторов $x, y \in \mathbb{R}^n$, где x_i – i-ая компонента вектора x в стандартном базисе в $\mathbb{R}^n$. Скалярное произведение порождает ℓ_2-норму в $\mathbb{R}^n$ в следующем виде $\|x\|_2 = \sqrt{\langle x, x\rangle}$. Обозначим ℓ_p-норму, как $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ для $p \in (1, \infty)$, а для $p = \infty$ используем $\|x\|_\infty = \max_{1\le i\le n} |x_i|$. Двойственная норма $\|\cdot\|_*$ для нормы $\|\cdot\|$ определяется как: $\|y\|_* = \max\{\langle x, y\rangle \mid \|x\| \le 1\}$. Для максимального и минимального положительных собственных значений положительно переопределенной матрицы $A \in \mathbb{R}^{n\times n}$ используется $\lambda_{\max}(A)$ и $\lambda^+_{\min}(A)$, а под $\chi(A) = {\lambda_{\max}(A)}/{\lambda^+_{\min}(A)}$ понимается число обусловленностей матрицы A. $\mathbb{E}[\cdot]$ – полное математическое ожидание, а $\mathbb{E}_\xi[\cdot]$ – условное математическое ожидание по случайной величине ξ. Приведение Кронеркера двух матриц $A \in \mathbb{R}^{m\times m}$ и $B \in \mathbb{R}^{n\times n}$ обозначается через $A \otimes B \in \mathbb{R}^{nm\times nm}$. Единичная матрица размера $n \times n$ определяется через I_n.

Поскольку все нормы эквивалентны в случае конечномерного пространства, то существуют такие константы C_1, C_2 и C_3, что для любого $x \in \mathbb{R}^n$

$$\|x\|_* \le C_1\|x\|_2, \quad \|x\|_2 \le C_2\|x\|_*, \quad \|x\| \le C_3\|x\|_2. \tag{2}$$

[1] Данная работа — это результат статей: [24, 25]

Например, если $\|\cdot\| = \|\cdot\|_2$, тогда $C_1 = C_2 = C_3 = 1$, или если $\|\cdot\| = \|\cdot\|_1$, тогда $\|\cdot\|_* = \|\cdot\|_\infty$, а $C_1 = 1$, $C_2 = C_3 = \sqrt{n}$.

Определение 1 (L-гладкость) *Функция g называется L-гладкой на $X \subseteq \mathbb{R}^n$ с константой $L > 0$ относительно нормы $\|\cdot\|$, если функция дифференцируемая и ее градиент является L-Липшецевым на X, т.е.*

$$\|\nabla g(x) - \nabla g(y)\|_* \leq L\|x - y\|, \quad \forall x, y \in X.$$

Можно показать, что из L-гладкости следует (смотри [26])

$$g(x) \leq g(y) + \langle \nabla g(y), x - y\rangle + \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in X. \tag{3}$$

Определение 2 (s-окрестность множества) *Для данного множества $X \subseteq \mathbb{R}^n$ и $s > 0$ s-окрестность X по норме $\|\cdot\|$ обозначим X_s, которое определяется следующим образом $X_s = \{z \in \mathbb{R}^n \mid \exists\, x \in X :\ \|y - x\| \leq s\}$.*

Определение 3 (Дивергенция Брегмана) *Пусть функция $\nu(x)$ является 1-сильно выпуклой по $\|\cdot\|$-норме и дифференцируемая на X функция. Тогда для любых двух точек $x, y \in X$ определим дивергенцию Брегмана $V(x, y)$ связанную с $\nu(x)$ как:*

$$V(x, y) = \nu(y) - \nu(x) - \langle \nabla\nu(x), y - x\rangle.$$

Заметим, что 1-сильно выпуклость $\nu(x)$ влечет

$$V(x, y) \geq \frac{1}{2}\|x - y\|^2. \tag{4}$$

Наконец, введем брегмановский диаметр множества X по дивергенции $V(x, y)$, как $D_{X,V} = \max\{\sqrt{2V(x, y)} \mid x, y \in X\}$. Принимая во внимание (4), можно заметить, что $D_{X,V}$ есть верхняя граница стандартного диаметра множества $D_X = \max\{\|x - y\| \mid x, y \in X\}$. Когда $V(x, y) = \frac{1}{2}\|x - y\|_2^2$, имеем $D_{X,V} = D_X$. Если $\|\cdot\| = \|\cdot\|_1$ есть ℓ_1-норма, то в случае когда X является вероятностным симплексом, т.е. $X = \{x \in \mathbb{R}^n_+ \mid \sum_{i=1}^n x_i = 1\}$, и расстояние определяется функцией энтропии $\nu(x)$, т.е. $\nu(x) = \sum_{i=1}^n x_i \ln x_i$, имеем, что $V(x, y)$ есть дивергенция Кульбака-Лейблера, т.е. $V(x, y) = \sum_{i=1}^n x_i \ln \frac{x_i}{y_i}$, и $D_{X,V} = \sqrt{2 \ln n}$ (смотри [27]).

2.2. Выпуклый случай

Рассматривается задача композитной оптимизации

$$\min_{x \in X} \Psi_0(x) = f(x) + g(x), \tag{5}$$

где $X \subseteq \mathbb{R}^n$ – компактное выпуклое множество с диаметром D_X в $\|\cdot\|$-норме, функция g выпуклая и L-гладкая на X, f дифференцируемая на X. Пусть имеется доступ к оракулу первого порядка для функции g, т.е. к градиенту $\nabla g(x)$, и к смещенному стохастическому оракулу нулевого порядка для f (смотри, например [28]), такой, что для данной точки x он возвращает зашумленное значение $\widetilde{f}(x)$:

$$\tilde{f}(x) = f(x, \xi) + \Delta(x), \tag{6}$$

где $\Delta(x)$ есть ограниченный шум неизвестной природы

$$|\Delta(x)| \leq \Delta, \tag{7}$$

а случайная величина ξ такая, что

$$\mathbb{E}[f(x, \xi)] = f(x). \tag{8}$$

Дополнительно предполагаем, что для любого $x \in X_s$ $(s \leq D_X)$

$$\|\nabla f(x, \xi)\|_2 \leq M(\xi), \quad \mathbb{E}[M^2(\xi)] = M^2. \tag{9}$$

Это предположение влечет, что для любого $x \in X_s$

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

и

$$\|\nabla f(x)\|_2 \leq M.$$

Используя такой оракул можно сконструировать аппроксимацию $\nabla f(x)$ в виде конечной разности (смотри [20, 21]):

$$\tilde{f}'_r(x) = \frac{n}{2r}(\tilde{f}(x + re) - \tilde{f}(x - re))e, \tag{10}$$

где e – случайный вектор равномерно распределенный на евклидовой сфере, а

$$r < sC_3 \tag{11}$$

есть параметр гладкости. Неравенство (11) гарантирует, что такая аппроксимация использует точку только из s-окрестности X, так как $\|re\| \leq rC_3$ (смотри (2)). Поэтому предполагаем, что всегда выполнено (11). Как и в [21], предполагается, что существует константа $p_* > 0$, что

$$\sqrt[4]{\mathbb{E}[\|e\|_*^4]} \quad \leq \quad p_*. \tag{12}$$

Например, когда $\|\cdot\| = \|\cdot\|_2$, имеем $p_* = 1$, а для случая, когда $\|\cdot\| = \|\cdot\|_1$, можно показать, что $p_* = O\left(\sqrt{\ln(n)/n}\right)$ (смотри Следствия 2 и 3 из [21]). Рассмотрим также сглаженную версию функции f:

$$F(x) = \mathbb{E}_e[f(x + re)], \tag{13}$$

которая является дифференцируемой X. Далее приведем некоторые полезные свойства функции $F(x)$.

Лемма 1 (смотри также Лемму 8 из [21]) *Пусть дифференцируемая функция f определенная на X_s, а также $\|\nabla f(x)\|_2 \leq M$ с некоторой константой $M > 0$. Тогда $F(x)$, определенная в (13), является выпуклой, дифференцируемой, и $F(x)$ удовлетворяет следующему условию*

$$\sup_{x \in X} |F(x) - f(x)| \quad \leq \quad rM, \tag{14}$$

$$\nabla F(x) \quad = \quad \mathbb{E}_e\left[\frac{n}{r} f(x + re)e\right], \tag{15}$$

$$\|\nabla F(x)\|_* \quad \leq \quad \tilde{c} p_* \sqrt{n} M, \tag{16}$$

где $\tilde{c}$ – некоторая константа, а p_* определена выше в (12).

Таким образом, $F(x)$ представляет собой хорошую аппроксимацию $f(x)$ для некоторого небольшого r. Следовательно, вместо того, чтобы решать задачу (5), можно сконцентрироваться на задаче

$$\min_{x \in X} \Psi(x) = F(x) + g(x) \tag{17}$$

с достаточно малым r. Следующая лемма говорит о полезных свойствах связи между $\nabla F(x)$ и $\tilde{f}'_r(x)$, определенного в (10).

Лемма 2 (модификация Леммы 10 из [21]) *Для $\widetilde{f}'_r(x)$, определенной в (10), справедливы следующие неравенства:*

$$\|\mathbb{E}[\tilde{f}'_r(x)] - \nabla F(x)\|_* \leq \frac{n \Delta p_*}{r}, \tag{18}$$

$$\mathbb{E}[\|\tilde{f}'_r(x)\|_*^2] \leq 2p_*^2\left(cnM^2 + \frac{n^2\Delta^2}{r^2}\right), \tag{19}$$

где c – некоторая положительная константа.

Таким образом, можно рассматривать $\tilde{f}'_r(x)$, как смещенный стохастический градиент $F(x)$ с ограниченным вторым моментом, и применить Sliding Algorithm из [1, 2] для решения задачи (17).

Algorithm 1 Zeroth-Order Sliding Algorithm (`zoSA`)

Вход: Начальная точка $x_0 \in X$ и максимальное число итераций N.

Пусть $\beta_k \in \mathcal{R}_{++}, \gamma_k \in \mathcal{R}_+$, и $T_k \in \mathbb{N}$, $k = 1, 2, \ldots$ определены, и положим $\overline{x}_0 = x_0$.

for $k = 1, 2, \ldots, N$ **do**

1. Положим $\underline{x}_k = (1-\gamma_k)\overline{x}_{k-1} + \gamma_k x_{k-1}$, и пусть $h_k(\cdot) \equiv l_g(\underline{x}_k, \cdot)$, где $l_g(\underline{x}_k, \cdot)$ из (22).
2. Положим
$$(x_k, \tilde{x}_k) = \texttt{PS}(h_k, x_{k-1}, \beta_k, T_k);$$
3. Положим $\overline{x}_k = (1-\gamma_k)\overline{x}_{k-1} + \gamma_k\tilde{x}_k$.

end for

Выход: $\overline{x}_N$.

`PS` (prox-sliding) процедура.

Процедура: $(x^+, \tilde{x}^+) = \texttt{PS}(h,\, x,\, \beta,\, T)$

Пусть параметры $p_t \in \mathbb{R}_{++}$ and $\theta_t \in [0, 1]$, $t = 1, \ldots$ определены. Пусть $u_0 = \tilde{u}_0 = x$.

for $t = 1, 2, \ldots, T$ **do**

$$u_t = \operatorname*{argmin}_{u \in X}\left\{h(u) + \langle \tilde{f}'_r(u_{t-1}), u\rangle + \beta V(x, u) + \beta p_t V(u_{t-1}, u)\right\}, \tag{20}$$
$$\tilde{u}_t = (1-\theta_t)\tilde{u}_{t-1} + \theta_t u_t. \tag{21}$$

end for

Пусть $x^+ = u_T$ и $\tilde{x}^+ = \tilde{u}_T$.

Конец процедуры:

В Algorithm 1 использовалась следующую функцию

$$l_g(x,y) = g(x) + \langle \nabla g(x), y - x \rangle. \tag{22}$$

На каждой итерации `PS` процедуры независимо генерируется новое направление e. Подчеркнем, что не нужно вычислять значения $F(x)$, что в общем случае требует численного вычисления интегралов по сфере. Напротив, метод требует знать только зашумленные значения f, определенные в (6).

Далее рассматривается анализ сходимости `zoSA`, который основан на анализе для метода Sliding Algorithm из [1, 2]. Следующая лемма анализирует процедуру `PS`.

Лемма 3 (модификация Предположения 8.3 из [2]) *Пусть* $\{p_t\}_{t\geq 1}$ *и* $\{\theta_t\}_{t\geq 1}$ *в процедуре* `PS` *удовлетворяют*

$$\begin{aligned} \theta_t &= \frac{P_{t-1} - P_t}{(1-P_t)P_{t-1}}, \\ P_t &= \begin{cases} 1 & t = 0, \\ p_t(1+p_t)^{-1}P_{t-1} & t \geq 1. \end{cases} \end{aligned} \tag{23}$$

Тогда для любого $t \geq 1$ *и* $u \in X$*:*

$$\begin{aligned} &\beta(1-P_t)^{-1}V(u_t,u) + [\Phi(\tilde{u}_t) - \Phi(u)] \\ &\leq \beta P_t(1-P_t)^{-1}V(u_0,u) + P_t(1-P_t)^{-1}\sum_{i=1}^{t}(p_iP_{i-1})^{-1}\left[\tfrac{(\tilde{M}+\|\delta_i\|_*)^2}{2\beta p_i} + \langle \delta_i, u - u_{i-1}\rangle\right], \end{aligned} \tag{24}$$

где

$$\Phi(u) = h(u) + F(u) + \beta V(x,u), \tag{25}$$

$$\delta_t = \tilde{f}'_r(u_{t-1}) - \nabla F(u_{t-1}). \tag{26}$$

$$\tilde{M} = c\sqrt{n}C_1M,$$

c *– некоторая положительная константа,* C_1 *определена в* (2)*.*

Используя приведенную выше лемму, получаем основной результат.

Теорема 1. *Предположим, что* $\{p_t\}_{t\geq 1}$*,* $\{\theta_t\}_{t\geq 1}$*,* $\{\beta_k\}_{k\geq 1}$*,* $\{\gamma_k\}_{k\geq 1}$ *в Algorithm 1 удовлетворяют* (23) *и*

$$\gamma_1 = 1, \qquad \beta_k - L\gamma_k \geq 0, \quad k \geq 1, \tag{27}$$

$$\frac{\gamma_k\beta_k}{\Gamma_k(1-P_{T_k})} \leq \frac{\gamma_{k-1}\beta_{k-1}}{\Gamma_{k-1}(1-P_{T_{k-1}})}, \quad k \geq 2. \tag{28}$$

Тогда

$$\begin{aligned}&\mathbb{E}[\Psi(\overline{x}_N) - \Psi(x^*)] \\ &\quad \leq \tfrac{\Gamma_N\beta_1}{1-P_{T_1}}V(x_0,u) + \Gamma_N \sum_{k=1}^{N}\sum_{i=1}^{T_k}\left[\tfrac{(\tilde{M}^2+\sigma^2)\gamma_k P_{T_k}}{\beta_k\Gamma_k(1-P_{T_k})p_i^2 P_{i-1}} + \tfrac{n\Delta D_X p_*}{r}\cdot\tfrac{\gamma_k P_{T_k}}{\Gamma_k(1-P_{T_k})p_i P_{i-1}}\right],\end{aligned} \tag{29}$$

где x^* – произвольная оптимальная точка (17), P_t определено в (23),

$$\Gamma_k = \begin{cases}1, & k=1,\\ (1-\gamma_k)\Gamma_{k-1}, & k>1\end{cases} \tag{30}$$

и

$$\sigma^2 = 4p_*^2\left(CnM^2 + \frac{n^2\Delta^2}{r^2}\right), \tag{31}$$

где C – некоторая положительная константа.

Рассмотрим конкретный выбор параметров и получим оценки на сходимость в более явном виде.

Следствие 1. *Пусть $\{p_t\}_{t\geq 1}$, $\{\theta_t\}_{t\geq 1}$ такие, что*

$$p_t = \frac{t}{2}, \quad \theta_t = \frac{2(t+1)}{t(t+3)}, \quad \forall t \geq 1, \tag{32}$$

N дано, а $\{\beta_k\}$, $\{\gamma_k\}$, $\{T_k\}$ есть

$$\beta_k = \frac{2L}{k}, \quad \gamma_k = \frac{2}{k+1}, \quad T_k = \frac{N(\tilde{M}^2+\sigma^2)k^2}{\tilde{D}L^2} \tag{33}$$

для константы $\tilde{D} = {}^{3D^2_{X,V}}\!/_4$. Тогда $\forall N \geq 1$

$$\mathbb{E}[\Psi(\overline{x}_N) - \Psi(x^*)] \leq \frac{12LD^2_{X,V}}{N(N+1)} + \frac{n\Delta D_X p_*}{r}. \tag{34}$$

Наконец, перенесем приведенный выше результат на исходную задачу (5).

Следствие 2. *В предположениях Следствия 1 следующие оценки справедливы для любого $N \geq 1$:*

$$\mathbb{E}[\Psi_0(\overline{x}_N) - \Psi_0(x^*)] \;\leq\; 2rM + \frac{12LD^2_{X,V}}{N(N+1)} + \frac{n\Delta D_X p_*}{r}. \tag{35}$$

Если взять

$$r \;=\; \Theta\left(\frac{\varepsilon}{M}\right), \quad \Delta = O\left(\frac{\varepsilon^2}{nMD_X\min\{p_*,1\}}\right) \tag{36}$$

и $\varepsilon = O\left(\sqrt{n}MD_X\right)$, $s = \Omega\left(\varepsilon/MC_3\right)$, *тогда количество вызовов оракулов* ∇g *и* $\tilde{f}'_r$, *которые делает Algorithm 1, чтобы найти* ε*-решение задачи* (5), *т.е. такие* $\overline{x}_N$, *что* $\mathbb{E}[\Psi_0(\overline{x}_N)] - \Psi_0(x^*) \leq \varepsilon$, *может быть ограничено*

$$O\left(\sqrt{\frac{LD_{X,V}^2}{\varepsilon}}\right), \tag{37}$$

$$O\left(\sqrt{\frac{LD_{X,V}^2}{\varepsilon}} + \frac{D_{X,V}^2 nM^2(C_1^2 + p_*^2)}{\varepsilon^2}\right). \tag{38}$$

Проанализируем полученные результаты. Прежде всего нас интересуют оценки (37) и (38). Для начала рассмотрим евклидов случай, т.е. $\|\cdot\| = \|\cdot\|_2$, $V(x,y) = \frac{1}{2}\|x-y\|_2^2$, $D_{X,V} = D_X$. В этом случае $p_* = C_1 = C_2 = C_3 = 1$ и оценка (38) для количества вызовов оракула (6) есть

$$O\left(\sqrt{\frac{LD_X^2}{\varepsilon}} + \frac{D_X^2 nM^2}{\varepsilon^2}\right),$$

а количество вызовов оракула $\nabla g(x)$ остается тем же. Это означает, что наш результат дает то же количество вызовов оракула первого порядка, что и в оригинальном Sliding Algorithm, в то время как число вызовов оракула нулевого порядка в n раз больше, чем в оригинальном методе первого порядка. В евклидовом случае наши оценки совпадают с известными оценками вызовов оракулов нулевого порядка (смотри [22]).

Далее рассмотрим случай, когда X есть вероятностный симплекс в $\mathbb{R}^n$. Как упоминалось ранее в разделе 2.1, в этой ситуации $D_{X,V} = \sqrt{2\ln n}$, $D_X = 2$, $p_* = O\left(\ln(n)/n\right)$, и $C_1 = 1$, $C_2 = C_3 = \sqrt{n}$. Тогда количество вызовов оракула $\nabla g(x)$ есть $O\left(\sqrt{(L\ln^2 n)/\varepsilon}\right)$. А для оракула $\tilde{f}'_r(x)$, имеем следующую оценку:

$$O\left(\sqrt{\frac{L\ln^2 n}{\varepsilon}} + \frac{M^2\ln^2 n}{\varepsilon^2}\right).$$

Понятно, что в этом случае имеем только полилогарифмическую зависимость от размерности n.

2.3. Сильно выпуклый случай

В этой части работы дополнительно предполагается, что функция g является μ-сильно выпуклой относительно дивергенции Брегмана $V(x,y)$, т.е. $\forall x, y \in X$

$$g(x) \geq g(y) + \langle \nabla g(y), x - y\rangle + \mu V(x,y).$$

Как и в оригинальной работе [1], здесь используется техника рестартов.

Algorithm 2 The Multi-phase Zeroth-Order Sliding Algorithm (`M-zoSA`)

Input: Начальную точку $y_0 \in X$ и максимальное число итерации N_0, начальное оценка параметра ρ_0 $(\Psi(y_0) - \Psi^* \leq \rho_0)$.

for $i = 1, 2, \ldots, I$ **do**

Запустить `zoSA` с $x_0 = y_{i-1}$, $N = N_0$, $\{p_t\}$ и $\{\theta_t\}$ из (32), $\{\beta_k\}$ и $\{\gamma_k\}$, $\{T_k\}$ из (33) с $\tilde{D} = \rho_0/\mu 2^i$, и пусть в y_i впишем выход этого шага.

end for

Выход: y_I.

Следующая теорема устанавливает основные результаты сходимости для `M-zoSA`.

Теорема 2. *Для* `M-zoSA` *с* $N_0 = 2\lceil\sqrt{5L/\mu}\rceil$ *имеет место следующее неравенство*

$$\mathbb{E}[\Psi(y_i) - \Psi(y^*)] \leq \frac{\rho_0}{2^i} + \frac{2n\Delta D_X p_*}{r}. \tag{39}$$

Следствие 3. *Для любого* $N \geq 1$ *процесс* `M-zoSA` *удовлетворяет*

$$\mathbb{E}[\Psi_0(y_i) - \Psi_0(y^*)] \leq 2rM + \frac{\rho_0}{2^i} + \frac{2n\Delta D_X p_*}{r}. \tag{40}$$

Из (40) *следует, что если*

$$r = \Theta\left(\frac{\varepsilon}{M}\right), \quad \Delta = O\left(\frac{\varepsilon^2}{nMD_X \min\{p_*, 1\}}\right) \tag{41}$$

и $\varepsilon = O\left(\sqrt{n}MD_X\right)$, $s = \Omega\left(\varepsilon/MC_3\right)$, *то количество вызовов оракулов* ∇g *и* $\tilde{f}'_r$, *которые делает Algorithm 2 для нахождения* ε*-решения задачи* (5) *может быть ограничено следующими соотношениями*

$$O\left(\sqrt{\frac{L}{\mu}}\log_2 \max\left[1, \rho_0/\varepsilon\right]\right), \tag{42}$$

$$O\left(\sqrt{\frac{L}{\mu}}\log_2 \max\left[1, \rho_0/\varepsilon\right] + \frac{nM^2(C_1^2 + p_*^2)}{\mu\varepsilon}\right). \tag{43}$$

3. От композитной оптимизации к выпуклой оптимизации с аффинными ограничениями и децентрализованной распределенной оптимизации

В этом разделе полученные результаты применяются к задачам выпуклой оптимизации с аффинными ограничениями, а затем к задаче децентрализованной рас-

пределенной оптимизации.

3.1. Выпуклая оптимизация с аффинными ограничениями

В качестве промежуточного шага между композитной оптимизацией (5) и децентрализованной распределенной оптимизации рассмотрим следующую проблему

$$\min_{Ax=0, x\in X} f(x), \tag{44}$$

где $A \succeq 0$, и $\mathrm{Ker} A \neq \{0\}$, а X является выпуклым и компактным множеством в $\mathbb{R}^n$ с диаметром D_X. Двойственная задача для (44) может быть записана в следующем виде

$$\begin{aligned} & \min_y \psi(y), \quad \text{где} \\ \varphi(y) &= \max_{x\in X} \{\langle y, x\rangle - f(x)\}, \\ \psi(y) &= \varphi(A^\top y) = \max_{x\in Q} \{\langle y, Ax\rangle - f(x)\} \\ &= \langle y, Ax(A^\top y)\rangle - f(x(A^\top y)) \\ &= \langle A^\top y, x(A^\top y)\rangle - f(x(A^\top y)), \end{aligned} \tag{45}$$

а $x(y) = \arg\max_{x\in X} \{\langle y, x\rangle - f(x)\}$. Решение задачи (45) наименьшее по ℓ_2-нормой обозначим y_*. Это норма $R_y = \|y_*\|_2$ может быть ограничена следующим образом [6]:

$$R_y^2 \leq \frac{\|\nabla f(x^*)\|_2^2}{\lambda_{\min}^+(A^\top A)}.$$

Как и в [11, 12, 29] рассмотрим следующую задачу

$$\min_{x\in X} F(x) = f(x) + \frac{R_y^2}{\varepsilon}\|Ax\|_2^2, \tag{46}$$

где ε – некоторое положительное число. Оказывается (смотри подробности в [12]), если имеется такой $\hat{x}$, что $F(\hat{x}) - \min_{x\in X} F(x) \leq \varepsilon$, тогда справедливо

$$f(\hat{x}) - \min_{Ax=0, x\in X} f(x) \leq \varepsilon, \quad \|A\hat{x}\|_2 \leq \frac{2\varepsilon}{R_y}.$$

Заметим, что этот результат можно обобщить следующим образом: если имеем $\hat{x}$, что $\mathbb{E}[F(\hat{x})] - \min_{x\in X} F(x) \leq \varepsilon$, тогда справедливо

$$\mathbb{E}[f(\hat{x})] - \min_{Ax=0, x\in X} f(x) \leq \varepsilon, \quad \sqrt{\mathbb{E}[\|A\hat{x}\|_2^2]} \leq \frac{2\varepsilon}{R_y}. \tag{47}$$

Далее рассмотрим проблемы (46), как композитную задачу (5) с $g(x) = {}^{R_y^2\|Ax\|_2^2}/_{\varepsilon}$. Предположим, что $\|\nabla f(x)\|_2 \leq M$ для любого $x \in X$, а для f имеется смещенный

стохастический оракул вида (6). Интересует ситуация, когда $\nabla g(x) = {}^{2R_y^2 A^\top Ax}/_{\varepsilon}$ может быть вычислено точно. Более того, легко заметить, что $g(x)$ является ${}^{2R_y^2 \lambda_{\max}(A^\top A)}/_{\varepsilon}$-гладкой по ℓ_2-норме. Применяя Следствие 2, получаем, что для того, чтобы получить точку $\hat{x}$, которая удовлетворяет (47), Algorithm 1 делает

$$O\left(\sqrt{\frac{\lambda_{\max}(A^\top A) R_y^2 D_X^2}{\varepsilon^2}}\right) \text{ вычислений } A^\top Ax$$

и

$$O\left(\sqrt{\frac{\lambda_{\max}(A^\top A) R_y^2 D_X^2}{\varepsilon^2}} + \frac{n D_X^2 M^2}{\varepsilon^2}\right)$$

запросов к оракулу $\widetilde{f}(x)$, так как $p_* = C_2 = C_1 = 1$ для евклидовой нормы. Как упомянуто в разделе 2, это граница зависит от n в стандартном виде.

3.2. Децентрализованная распределенная оптимизация

Вернемся к проблеме (1), как и в [7], можем переписать ее в следующем виде:

$$\min_{\substack{x_1 = \ldots = x_m \\ x_1, \ldots, x_m \in X}} f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} f_i(x_i), \tag{48}$$

где $\mathbf{x}^\top = (x_1^\top, \ldots, x_m^\top)^\top \in \mathbb{R}^{nm}$. Напомним, что рассматривается ситуацию, когда f_i хранится на i-ом узле. В этом случае можно интерпретировать x_i из (48), как локальную переменную i-го узла, а $x_1 = \ldots = x_n$ рассматривать в качестве условия консенсуса для сети. Стандартный трюк [7–9, 13] работать с этим условием – переписать его, используя понятие матрицы Лапласа. Матрица Лапласа $\overline{W} = \|\overline{W}_{ij}\|_{i,j=1,1}^{m,m} \in \mathbb{R}^{m \times m}$ графа G с множеством вершин V, $|V| = m$ и множеством ребер V определяется, как:

$$\overline{W}_{ij} = \begin{cases} -1, & \text{если } (i,j) \in E, \\ \deg(i), & \text{если } i = j, \\ 0 & \text{иначе,} \end{cases}$$

где $\deg(i)$ есть степень i-ой вершины. В этой работе будет рассматриваться только сети с одной компонентой связности. В этом случае $\overline{W}$ имеет уникальный собственный вектор $1_m = (1, \ldots, 1)^\top \in \mathbb{R}^m$ с собственный значением 0. Используя это, можно показать, что для любого вектор $a = (a_1, \ldots, a_m)^\top \in \mathbb{R}^m$ справедливо следующее выражение:

$$a_1 = \ldots = a_m \iff Wa = 0. \tag{49}$$

Используя произведение Кронекера, введем $W = \overline{W} \otimes I_n$. Полученная матрица также называется матрицей Лапласа для простоты, несложно обобщить (49) на n-мерный случай:

$$x_1 = \ldots = x_m \iff W\mathbf{x} = 0$$

и

$$x_1 = \ldots = x_m \iff \sqrt{W}\mathbf{x} = 0.$$

То есть вместо задачи (48) можно рассмотреть эквивалентную задачу

$$\min_{\substack{\sqrt{W}\mathbf{x}=0,\\ x_1,\ldots,x_m \in X}} f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(x_i). \tag{50}$$

Далее нам нужно определить параметры f, используя локальные параметры f_i. Предположим, что каждая f_i удовлетворяет следующим условиям: $\|f_i(x_i)\|_2 \leq M$ для любого $x_i \in X$, все f_i являются выпуклыми. Тогда можно показать (смотри [12]), что $\|\nabla f(\mathbf{x})\|_2 \leq M/\sqrt{m}$ на множестве таких $\mathbf{x}$ для $x_1, \ldots, x_m \in X$, $D^2_{X^m} = mD^2_X$ и $R^2_{\mathbf{y}} = \|\mathbf{y}_*\|^2_2 \leq M^2/m\lambda^+_{\min}(W)$.

Теперь применим результаты, полученные в разделе 3.1 для задачи (50). Действительно, эту проблему можно рассматривать, как (50) с $A = \sqrt{W}$. Принимая это во внимание, замечаем, что одно вычисление $A^\top Ax$ соответствует вычислению Wx, которое можно вычислить в течение одного раунда связи в сети с матрицей Лапласа W. Это простое наблюдение означает, что для получения такой точки $\hat{\mathbf{x}}$, которая удовлетворяет (47) с $\hat{x} = \hat{\mathbf{x}}$, $A := \sqrt{W}$, $X := X^n$, $R_y := R_{\mathbf{y}}$, Algorithm 1 требует

$$O\left(\sqrt{\frac{\chi(W)M^2D^2_X}{\varepsilon^2}}\right) \text{ раундов связи}$$

и

$$O\left(\sqrt{\frac{\chi(W)M^2D^2_X}{\varepsilon^2}} + \frac{nD^2_XM^2}{\varepsilon^2}\right)$$

вызовов $\widetilde{f}(x)$ на каждом устройстве, так как $p_* = 1$ для евклидовова случая. Оценка раундов связи соответствует нижней границе [8, 9], и мы считаем, что в наших предположениях полученная оценка для количества вычислений оракула нулевого порядка на каждом устройстве является оптимальной с точностью до полилогарифмических факторов в классе методов с оптимальным числом раундов связи (смотри [11, 12]).

4. Анализ допустимого множества

Как уже было сказано выше, в работах (см. [21, 23]), где вместо «честного» градиента используется оракул нулевого (10), важно, что функция определена не только на допустимом множестве, но и в некоторой его окрестности. Это связано с тем, что для любой точки x, принадлежащей множеству, точка $x + re$ может находиться за пределами множества.

Но в некоторых случаях нельзя сделать такое предположение. Функция и значения x могут иметь реальную физическую интерпретацию. Например, в случае вероятностного симплекса значения x являются распределением ресурсов или действий. Сумма вероятностей не может быть отрицательной или быть больше 1. Более того, из-за реализации или по другим причинам может иметься только оракул, который четко определен на допустимом множестве и больше нигде.

В этом разделе излагается подход к решению озвученной выше проблемы и то, как качество решения меняется от этого.

Подход можно кратко описать следующим образом:

- Уменьшить исходное множество X в $(1-\alpha)$ раз и рассмотреть «сжатую» версию X^α. Обратим внимание, что параметр α не должен быть слишком маленьким, так как в противном случае параметр r должен быть еще меньше. Но также нельзя брать большие α, потому что при этом слишком сильно сжимается множество и можно получить далеко не оптимальное решение. Это означает, что точность решения ε ограничивает α: $\alpha \leq h(\varepsilon)$, в свою очередь, α ограничивает r: $r \leq g(\alpha)$.

- Генерировать случайное направление e, так, что для любого $x \in X^\alpha$ следует $x + re \in X$.

- Решать задачу на «сжатом» множестве с точностью $\varepsilon/2$. Параметр α должен быть выбран так, чтобы мы нашли решение ε исходной задачи.

На практике это может быть реализовано следующим образом: 1) как описано в предыдущем абзаце, или 2) работать с исходным набором X, но если $x_k + re$ находится вне X , то спроецировать x_k на множество X^α. Далее приводим теоретический анализ только для метода, который всегда работает на X^α.

Далее анализируются случаи разных множеств. Общая схема анализа:

- Ввести способ «сжать» множество.
- Предложить стратегию генерации e случайного направления.
- Оцените минимальное расстояние между X^α и X в ℓ_2-норме. Это и есть ограничение на r, так как $\|e\|_2 = 1$.
- Оценить параметр α так, чтобы $\varepsilon/2$ – решение «сжатой» задачи не отличалось более чем на $\varepsilon/2$ от ε-решения исходной задачи.

Первый случай – **вероятностный симплекс**:

$$\triangle_n = \left\{ \sum_{i=1}^{n} x_i = 1, \quad x_i \geq 0, \quad i \in 1 \dots n \right\}.$$

Рассмотрим гиперплоскость

$$\mathcal{H} = \left\{ \sum_{i=1}^{n} x_i = 1 \right\},$$

на которой лежит симплекс. Обратите внимание, что если e выбирается так, что оно лежит на $\mathcal{H}$, тогда для любого x, лежащего на гиперплоскости, $x + re$ тоже будет лежать на ней. Поэтому генерируем случайное направление e на гиперплоскости. Обратите внимание, что $\mathcal{H}$ является подпространством $\mathbb{R}^n$ размерности $\dim\mathcal{H} = n-1$. Можно проверить, что множество векторов из $\mathbf{R}^n$

$$\mathbf{v} = \begin{pmatrix} \mathbf{v}_1 = 1/\sqrt{2}(1, -1, 0, 0, \dots 0), \\ \mathbf{v}_2 = 1/\sqrt{6}(1, 1, -2, 0, \dots 0), \\ \mathbf{v}_3 = 1/\sqrt{12}(1, 1, 1, -3, \dots 0), \\ \dots \\ \mathbf{v}_k = 1/\sqrt{k+k^2}(1, \dots 1, -k, \dots, 0), \\ \dots \\ \mathbf{v}_{n-1} = 1/\sqrt{n-1+(n-1)^2}(1, \dots, 1, -n+1) \end{pmatrix},$$

является ортонормированным базисом $\mathcal{H}$. Поэтому, генерируя равномерно вектор $\tilde{\mathbf{e}}$ на евклидовой сфере $\mathcal{RS}_2^{n-1}(1)$ и вычисляя $\mathbf{e}$ по следующей формуле:

$$e = \tilde{e}_1\mathbf{v}_1 + \tilde{e}_2\mathbf{v}_2 + \ldots + \tilde{e}_k\mathbf{v}_k + \ldots \tilde{e}_{n-1}\mathbf{v}_{n-1}, \tag{51}$$

мы получаем, что требуется. С таким вектором e мы всегда остаемся на гиперплоскости, но можем выйти за пределы симплекса. Это происходит тогда и только тогда, когда для некоторого i $x_i + re_i < 0$. Чтобы избежать этого, рассмотрим «сжатый» симплекс для некоторой положительной константы α:

$$\triangle_n^\alpha = \left\{ \sum_{i=1}^n x_i = 1, \quad x_i \geq \alpha, \quad i \in 1 \dots n \right\}.$$

Можно заметить, что для любого $x \in \triangle_n^\alpha$, для любых e из (51) и $r < \alpha$ следует, что $x + re \in \triangle_n$, потому что $|e_i| \leq 1$, а значит $x_i + re_i \geq \alpha - r \geq 0$. Можно заметить, что для любого $x \in \triangle_n^\alpha$, для любого e из (51) и $r < \alpha$ следует, что $x + re \in \triangle_n$, потому что $|e_i| \leq 1$, а значит $x_i + re_i \geq \alpha - r \geq 0$.

Последний вопрос, который необходимо обсудить, – это точность решения, которое получается на «сжатом» множестве. Рассмотрим следующую лемму:

Лемма 4. *Пусть функция $f(x)$ является M-гладкой относительно $\|\cdot\|_2$. Рассмотрим задачу минимизации функции $f(x)$, но не на исходном множестве X, а на «сжатом» множестве X_α. Пусть найдено x_k – решение с точностью $\varepsilon/2$ по $f(x)$. Тогда это есть $(\varepsilon/2 + tM)$-решение исходной задачи, где*

$$t = \max_{x \in X} \left\| x - \operatorname*{argmin}_{\hat{x} \in X^\alpha} \|x - \hat{x}\|_2 \right\|_2.$$

На самом деле, нет необходимости искать ближайшую точку к каждому x и находить t. Достаточно найти "довольно" близкую и оценить верхнюю границу t. Поэтому остается найти правило, которое каждой точке x из X ставит в соответствие точку $\hat{x}$ из X_α и оценить максимальное расстояние $\max_X \|\hat{x} - x\|_2$. Для точки симплекса рассмотрим следующее правило:

$$\hat{x}_i = \frac{(x_i + 2\alpha)}{(1 + 2\alpha n)}, \qquad i = 1, \dots n. \tag{52}$$

Легко видеть, что для любого $\alpha \leq 1/2n$:

$$\sum_{i=1}^n \hat{x}_i = 1, \qquad \hat{x}_i \leq \alpha, \qquad i = 1, \dots n.$$

Это означает, что $\hat{x} \in X_\alpha$. Расстояние $\|\hat{x} - x\|_2$:

$$\|\hat{x} - x\|_2 = \sqrt{\sum_{i=1}^n (\hat{x}_i - x_i)^2} = \frac{2\alpha n}{1 + 2\alpha n} \sqrt{\sum_{i=1}^n \left(\frac{1}{n} - x_i \right)^2}.$$

$\sqrt{\sum\limits_{i=1}^{n}\left(\frac{1}{n}-x_i\right)^2}$ есть просто расстояние до цента симплекса. Его можно оценить сверху радиусом описанной сферы $R=\sqrt{\frac{n-1}{n}}\leq 1$. Тогда

$$\|\hat{x}-x\|_2 \leq \frac{2\alpha n}{1+2\alpha n} \leq 2\alpha n. \tag{53}$$

(53) вместе с Леммой 4 дает, что $f(x_k)-f(x^*)\leq \frac{\varepsilon}{2}+2\alpha n M$. Взяв $\alpha = \varepsilon/4nM$ (или меньше), имеем ε-решение исходной задачи. Откуда $r\leq\alpha=\varepsilon/4nM$.

Второй случай – **положительный ортант**:

$$\perp_n = \{x_i \geq 0, \quad i \in 1\dots n\}.$$

Предлагается следующее «сжатое» множество:

$$\perp_n^\alpha = \{y_i \geq \alpha, \quad i \in 1\dots n\}.$$

Можно заметить, что для любого i минимум выражения $y_i + re_i$ есть $\alpha - r$, так как $e_i \geq -1$ и $y_i \geq \alpha$. Следовательно необходимо, чтобы $\alpha - r \geq 0$. Это означает, что для любого $e \in \mathcal{RS}_2^n(1)$, для вектора $y+re$ справедливо следующее выражение:

$$y_i + re_i \geq 0, \quad i \in 1\dots n.$$

Тогда найдем t в Лемме 4 для ортанта. Пусть для любого $x \in \perp_n$ определим $\hat{x}$ следующим образом:

$$\hat{x}_i = \begin{cases} \alpha, & x_i < \alpha, \\ x_i, & x_i \geq \alpha, \end{cases} \qquad i = 1,\dots n. \tag{54}$$

Заметим, что $\hat{x}_i \in \perp_n^\alpha$ и

$$\|\hat{x}-x\|_2 = \sqrt{\sum_{i=1}^{n}(\hat{x}_i - x_i)^2} \leq \sqrt{\sum_{i=1}^{n}\alpha^2} = \alpha\sqrt{n}. \tag{55}$$

По Лемме 4 имеем, что $f(x_k)-f(x^*)\leq\frac{\varepsilon}{2}+\alpha\sqrt{n}M$. Тогда с $\alpha=\varepsilon/2\sqrt{n}M$ (или меньше), мы найдем ε-решение исходной задачи. Откуда $r\leq\alpha=\varepsilon/2\sqrt{n}M$.

Третий вид множества – **шар в p-норме** для $p\in[1;2]$:

$$\mathcal{B}_p^n(a,R) = \{\|x-a\|_p \leq R\},$$

где a – центр шара, R – его радиус. Предлагается рассматривать «сжатый» шар вида $\mathcal{B}_p^n(a, R(1-\alpha))$.

Лемма 5. *Рассмотрим две концентрические сферы в p норме, где* $p \in [1;2]$, $\alpha \in (0;1)$*:*

$$\mathcal{S}_p^n(a,R) = \{\|x-a\|_p = R\}, \quad \mathcal{S}_p^n(a,R(1-\alpha)) = \{\|y-a\|_p = R(1-\alpha)\}.$$

Тогда минимальное расстояние между двумя сферами есть

$$m = \frac{\alpha R}{n^{1/p-1/2}}.$$

Используя эту лемму, можно заметить, что для любого $x \in \mathcal{B}_n^{\alpha}(a, R(1-\alpha))$, $r \leq {\alpha R}/{n^{1/p-1/2}}$ и для любого $e \in \mathcal{RS}_2^n(1)$, $x + re \in \mathcal{B}_n(a,R)$.

Тогда найдем t в Лемме 4 для шара. Пусть для любого x определим $\hat{x}$ в следующем виде:

$$\hat{x}_i = a + (1-\alpha)(x_i - a), \qquad i = 1, \dots n. \tag{56}$$

Можно видеть, что $\hat{x}_i$ принадлежит «сжатому» шару и

$$\|\hat{x}-x\|_2 = \sqrt{\sum_{i=1}^n (\hat{x}_i - x_i)^2} = \sqrt{\sum_{i=1}^n (\alpha(x_i-a))^2} = \alpha\sqrt{\sum_{i=1}^n (x_i-a)^2} \leq \alpha \sum_{i=1}^n |x_i - a|.$$

По неравенству Гельдера:

$$\|\hat{x}-x\|_2 \leq \alpha \sum_{i=1}^n |x_i - a| \leq \alpha n^{\frac{1}{q}} \left(\sum_{i=1}^n |x_i-a|^p\right)^{\frac{1}{p}} = \alpha n^{\frac{1}{q}} R.$$

По Лемме 4 имеем, что $f(x_k) - f(x^*) \leq \frac{\varepsilon}{2} + \alpha n^{1/q} RM$. Тогда при $\alpha = {\varepsilon}/{2n^{1/q}RM}$ (или меньше), мы найдем ε-решение исходной задачи. Откуда $r \leq {\alpha R}/{n^{1/p-1/2}} = {\varepsilon}/{2M\sqrt{n}}$.

Обобщим результаты этой части в Таблице 1.

Множество	α "сжатого" множества	Граница τ	$\mathbf{e}$
вероятностный симплекс	$\frac{\varepsilon}{4nM}$	$\frac{\varepsilon}{4nM}$	see (51)
положительный ортант	$\frac{\varepsilon}{2\sqrt{n}M}$	$\frac{\varepsilon}{2\sqrt{n}M}$	$\mathcal{RS}_2^n(1)$
шар в p-норме	$\frac{\varepsilon}{2n^{1/q}RM}$	$\frac{\varepsilon}{2\sqrt{n}M}$	$\mathcal{RS}_2^n(1)$

Таблица 1. Результаты Раздела 4

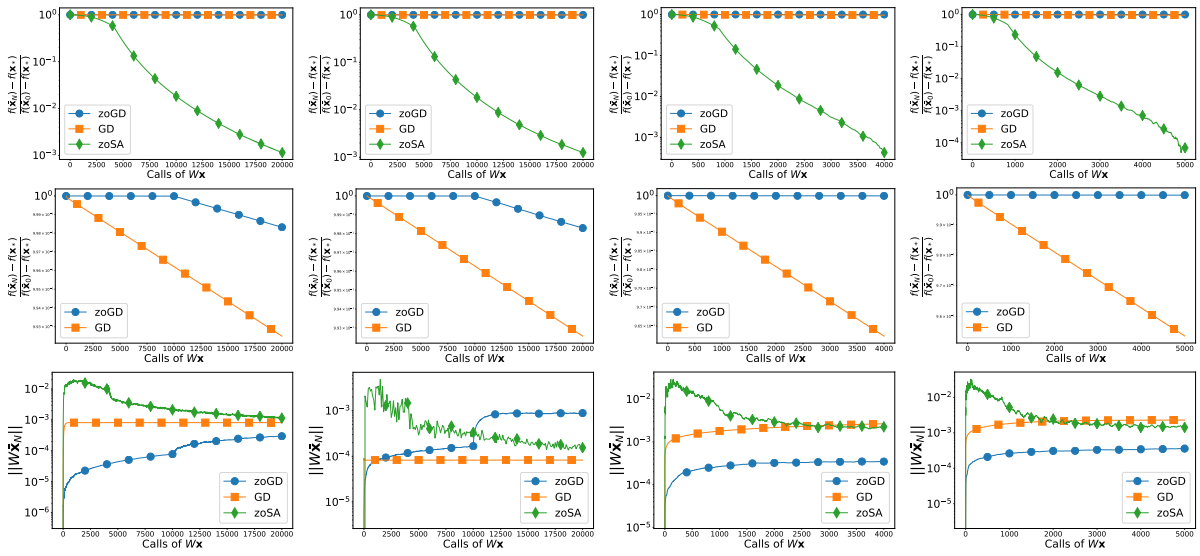
Можно заметить, что в (36) r не зависит от n. Исходя из результатов, полученных в этом разделе, нужно учесть зависимость от n. В Таблице 2, приведены новые ограничения на r и Δ.

Вторая колонка Таблицы 2 означает, определена ли функция в некоторой окрестности множества или нет.

Множество	Окр-сть?	τ	Δ
вероятностный симплекс	✓	$\Theta\left(\frac{\varepsilon}{M}\right)$	$O\left(\frac{\varepsilon^2}{nMD_X \min\{p_*,1\}}\right)$
	✗	$\Theta\left(\frac{\varepsilon}{Mn}\right)$ and $\leq \frac{\varepsilon}{4nM}$	$O\left(\frac{\varepsilon^2}{n^2MD_X \min\{p_*,1\}}\right)$
положительный ортант	✓	$\Theta\left(\frac{\varepsilon}{M}\right)$	$O\left(\frac{\varepsilon^2}{nMD_X \min\{p_*,1\}}\right)$
	✗	$\Theta\left(\frac{\varepsilon}{M\sqrt{n}}\right)$ and $\leq \frac{\varepsilon}{\sqrt{8n}M}$	$O\left(\frac{\varepsilon^2}{\sqrt{n^3}MD_X \min\{p_*,1\}}\right)$
шар в p-норме	✓	$\Theta\left(\frac{\varepsilon}{M}\right)$	$O\left(\frac{\varepsilon^2}{nMD_X \min\{p_*,1\}}\right)$
	✗	$\Theta\left(\frac{\varepsilon}{M\sqrt{n}}\right)$ and $\leq \frac{\varepsilon}{\sqrt{8n}M}$	$O\left(\frac{\varepsilon^2}{\sqrt{n^3}MD_X \min\{p_*,1\}}\right)$

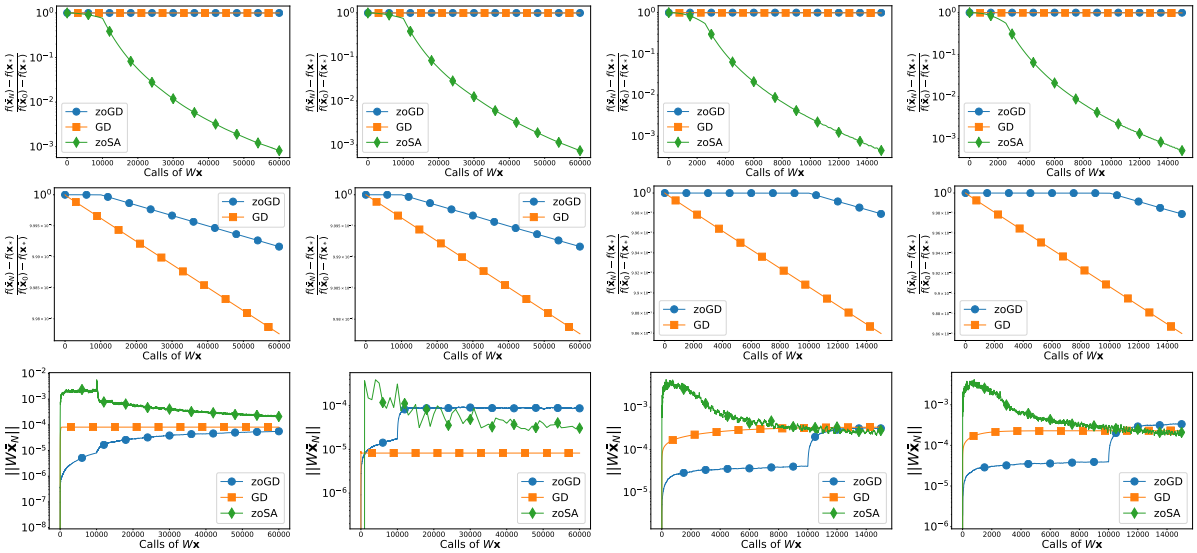
Таблица 2. r и Δ в Следствиях 2,3 для различных случаев

5. Численные эксперименты



(a) star (b) complete graph (c) chain d) cycle

Рис. 1. `zoSA`, `GD` and `zoGD` для решения задачи (57) с $R = 10^2$ для различных структур сети. Первые два ряда показывают, как изменяется относительное расхождение с решением для методов во время их работы, а последний ряд показывает эволюцию $W\overline{\mathbf{x}}_N$.



(a) star (b) complete graph (c) chain (d) cycle

Рис. 2. `zoSA`, `GD` and `zoGD` для задачи (57) с $R = 10^3$ для различных структур сети. Первые два ряда показывают, как изменяется относительное расхождение с решением для методов во время их работы, а последний ряд показывает эволюцию $W\overline{\mathbf{x}}_N$.

В экспериментах использовалось следующее оборудование: компьютер с 4 ядрами, каждое модели Intel(R) Core(TM) i7-9750H CPU @ 2.60 GHz. Сравниваются `zoSA`, зеркальный спуск [30] и зеркальный спуск нулевого порядка [23] для евклидового случая, т.е. градиентный спуск (`GD`) и его версию нулевого порядка (`zoGD`). Как уже упоминалось ранее, `zoSA` это первый метод, который для задачи (5) использует оракул первого порядка для гладкой части g и оракул нулевого порядка для негладкой части f. Поэтому стоит сравнивать `zoSA` с `GD` и `zoGD`, так как это современные и актуальные методы первого и нулевого порядка для выпуклых негладких задач.

5.1. Распределенное вычисление геометрической медианы

Рассмотрим проблему поиска геометрической медианы [31, 32] из m векторов $b_1, \ldots, b_m \in \mathbb{R}^n$:

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{m} \sum_{i=1}^{m} \|x - b_i\|_2.$$

Согласно разделу 3, рассмотрим следующую задачу:

$$\min_{\mathbf{x} \in \mathbb{R}^{nm}} F(\mathbf{x}) = \underbrace{\frac{1}{m} \sum_{i=1}^{m} \overbrace{\|x_i - b_i\|_2}^{f_i(x_i)}}_{f(\mathbf{x})} + \underbrace{R\|\sqrt{W}\mathbf{x}\|_2^2}_{g(\mathbf{x})}. \tag{57}$$

Как упоминалось ранее, если $R = {}^{R_{\mathbf{y}}^2}/_{\mathbf{e}}$, то $F(\overline{\mathbf{x}}) - \min_{\mathbf{x}\in\mathbb{R}^{nm}} F(\mathbf{x}) \leq \mathbf{e}$, то $f(\overline{\mathbf{x}}) - \min_{\sqrt{W}\mathbf{x}=0} f(\mathbf{x}) \leq \mathbf{e}$ и $\|\sqrt{W}\overline{\mathbf{x}}\|_2 \leq {}^{2\mathbf{e}}/_{R_{\mathbf{y}}}$. Однако на практике можно использовать разные варианты R, если с помощью такого выбора можно быстрее получить точку $\overline{\mathbf{x}}$ такую, что $\|\sqrt{W}\overline{\mathbf{x}}\|_2$ достаточно мало. Пробовались различные R, но лучшие результаты были получены для $R = 10^2$ и $R = 10^3$.

В наших экспериментах эмулируется работа децентрализованной распределенной системы с заданной матрицей Лапласа W на одной машине, чтобы продемонстрировать работу `zoSA` для задачи децентрализованной распределенной оптимизации. То есть мы храним $\mathbf{x}$ как составной вектор и считаем число вычислений $W\mathbf{x}$, так как это соответствует количеству коммуникационных раундов в распределенной системе. Во многих реальных распределенных сетях связь является узким местом, поэтому количество раундов связи в некоторой степени отражает время выполнения метода.

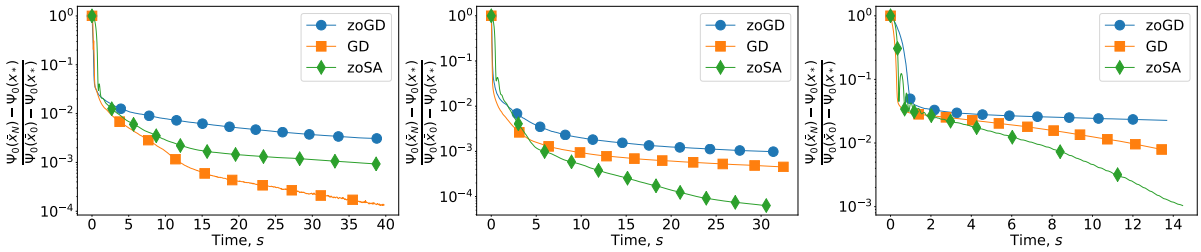
Тестировались `zoSA`, `GD` and `zoGD` на задаче (57) с $n = 10$ и $m = 100$ для стандартных структур сети: звезда, цикл, цепь, и полный граф. Векторы $b_1, \ldots, b_m$ генерировались независимо из нормального распределения $\mathcal{N}(1, 2I_n)$ со средним $1 = (1, \ldots, 1)^\top$ и ковариационной матрицей $2I_n$. Результаты представлены на Рис. 1 и 2. Заметим, что в этих экспериментах `zoSA` превосходит даже `GD`, который является методом первого порядка.

5.2. Логистическая регрессия с регуляризацией лассо

Далее рассматривается задача логистической регрессии с лассо регуляризацией для бинарной классификации:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \Psi_0(x) &= \overbrace{l_1\|x\|_1}^{f(x)} + g(x) \\ g(x) &= \frac{1}{m}\sum_{i=1}^{m} \log\left(1 + \exp\left(-y_i \cdot (Ax)_i\right)\right). \end{aligned} \tag{58}$$

Здесь $A \in \mathbb{R}^{m\times n}$ – матрица признаков объектов, $y_1, \ldots, y_m \in \{-1, 1\}$ – метки этих объектов, m – размер датасета, а $x \in \mathbb{R}^n$ – вектор весов. `zoSA`, `GD` и `zoGD` сравнивались на датасетах `mushrooms` ($m = 8124$, $n = 112$) с $l_1 = 10^{-3}$, `a5a` ($m = 6414$, $n = 123$) с $l_1 = 10^{-4}$ и `german.numer` ($m = 1000$, $n = 24$) с $l_1 = 10^{-4}$ [33], смотри Рис. 3. В первом случае для метода `zoSA` получилось сходимость, которая лучше, чем у `zoGD` и хуже, чем у `GD`, что кажется разумным для метода, который использует смешанный



(a) `mushrooms`, $l_1 = 10^{-3}$ (b) `a5a`, $l_1 = 10^{-4}$ (c) `german.numer`, $l_1 = 10^{-4}$

Рис. 3. `zoSA`, `GD` and `zoGD` для задачи (58) на различных датасетах и с различными параметрами l_1.

оракул. Тем не менее, наш метод превосходит даже `GD` на втором и третьем датасетах. Здесь нет противоречия: `zoSA` основан на Sliding Algorithm, который имеет лучшие теоретические оценки сходимости, чем `GD`, а `zoSA` имеет ту же сложность, что и Sliding Algorithm с точки зрения количества подсчетов $\nabla g(x)$.

5.3. Минимизация функции Нестерова с регуляризацией лассо

В этом разделе рассмотрим следующую задачу:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \Psi_0(x) &= \overbrace{l_1 \|x\|_1}^{f(x)} + g(x) \\ g(x) &= \frac{L}{8}\left(x_1^2 + \sum_{i=1}^{n-1}(x_i - x_{i+1})^2 + x_n^2\right) - \frac{L x_1}{4}. \end{aligned} \tag{59}$$

Здесь $g(x)$ выпуклая и L-гладкая функция, которая является одной из "наихудших" функций для методов первого порядка в классе выпуклых и L-гладких функций [26], а $f(x)$ имеет ограниченный градиент. Мы сравниваем `zoSA`, `GD` и `zoGD` на этой задаче с параметрами $L = 4$ и $l_1 = 10^{-3}$. Результаты представлены на Рис. 4. Естественно, `zoSA` превосходит `zoGD`, поскольку `zoSA` использует оракула первого порядка для гладкой части, тогда как `zoGD` использует только информацию нулевого порядка о $g(x)$. В то же время наш метод уступает `zoGD`, и в некоторой степени также ожидается: оракул первого порядка для $f(x)$ дает больше информации о направлении спуска, чем получает `zoSA` через оракула нулевого порядка.

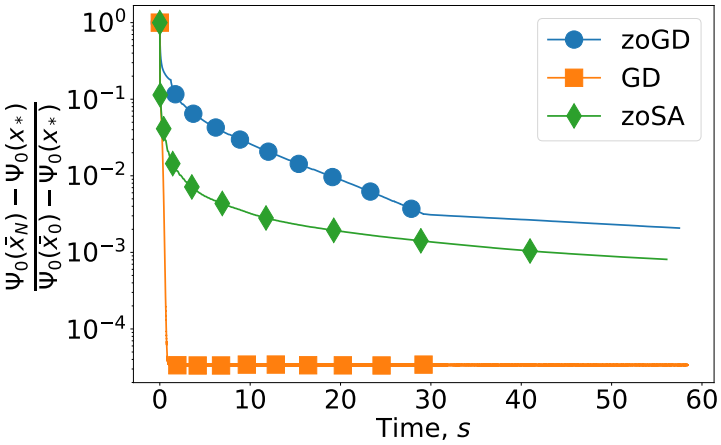


Рис. 4. `zoSA`, `GD` и `zoGD` для задачи (59) с $l_1 = 10^{-3}$ и $L = 4$.

6. Базовые факты

Простая верхняя оценка для квадрата суммы. Для произвольного натурального числа $n \geq 1$ и произвольного набора положительных чисел $a_1, \ldots, a_n$ справедливо

$$\left(\sum_{i=1}^{m} a_i\right)^2 \leq m \sum_{i=1}^{m} a_i^2 \tag{60}$$

Неравенство Гельдера. Для произвольных $x, y \in \mathbb{R}^n$ имеет место следующее неравенство

$$\langle x, y \rangle \leq \|x\|_* \cdot \|y\| \tag{61}$$

Неравенство Коши-Буняковского-Шварца. Пусть ξ и η – вещественные случайные величины, такие, что $\mathbb{E}[\xi^2] < \infty$ и $\mathbb{E}[\eta^2] < \infty$. Тогда

$$\mathbb{E}[\xi\eta] \leq \sqrt{\mathbb{E}[\xi^2]\mathbb{E}[\eta^2]}. \tag{62}$$

7. Вспомогательные результаты

Лемма 6 (Lemma 9 from [21]) *Для любой L -липшицева относительно ℓ_2-нормы функции g, справедливо*

$$\sqrt{\mathbb{E}[(g(e) - \mathbb{E}g(e))^4]} \leq c\frac{L^2}{n},$$

где e равномерно распределено на евклидовой единичной сфере, а c – некоторая константа.

Лемма 7 (Лемма 3.5 из [2]) *Пусть дана выпуклая функция $p : X \to \mathbb{R}$, точки $\tilde{x}, \tilde{y} \in X$ и константы $\mu_1, \mu_2 \geq 0$. Пусть $\nu : X \to \mathbb{R}$ дифференцируемая выпуклая функция, и $V(x, z)$:*

$$V(x, z) = \nu(z) - [\nu(x) + \nabla\nu(x)^\top (z - x)].$$

Если

$$\tilde{u} = \operatorname*{argmin}_{u \in X} \{p(u) + \mu_1 V(\tilde{x}, u) + \mu_2 V(\tilde{y}, u)\},$$

то для любого $u \in X$, справедливо

$$p(\tilde{u}) + \mu_1 V(\tilde{x}, \tilde{u}) + \mu_2 V(\tilde{y}, \tilde{u}) \leq p(u) + \mu_1 V(\tilde{x}, u) + \mu_2 V(\tilde{y}, u) - (\mu_1 + \mu_2) V(\tilde{u}, u).$$

Лемма 8 (Лемма 3.17 из [2]) *Пусть даны $w_k \in (0; 1]$, $k = 1, 2, \ldots$. Определим также*

$$W_k = \begin{cases} 1, & k = 1, \\ (1 - w_k) W_{k-1}, & k > 1. \end{cases}$$

Пусть $W_k > 0$ для любого $k > 1$, а последовательность $\{\delta_k\}_{k \geq 0}$ удовлетворяет

$$\delta_k \leq (1 - w_k)\delta_{k-1} + B_k, \quad k = 1, 2, \ldots$$

для некоторых положительных констант $\{B_k\}_{k \geq 0}$. Тогда справедливо

$$\delta_k \leq W_k (1 - w_1)\delta_0 + W_k \sum_{i=1}^{k} \frac{B_i}{W_i}.$$

8. Пропущенные доказательства из Раздела 2.2

8.1. Техническая лемма

Лемма 9. *Предположим, что для дифференцируемой функции f, определенной на замкнутом и выпуклом множестве X, существует такое M, что*

$$\|\nabla f(x)\|_2 \leq M \quad \forall x \in X. \tag{63}$$

Тогда

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + 2MC_1 \|x - y\|, \quad \forall x, y \in X.$$

Доказательство Леммы 9. Для произвольных точек $x, y \in X$ имеем

$$\begin{aligned} f(x) &= f(y) + \int_0^1 \langle \nabla f(y + \tau(x-y)), x-y\rangle d\tau \\ &= f(y) + \langle \nabla f(y), x-y\rangle + \int_0^1 \langle \nabla f(y+\tau(x-y)) - \nabla f(y), x-y\rangle d\tau \\ &\overset{(61)}{\leq} f(y) + \langle \nabla f(y), x-y\rangle + \int_0^1 \|\nabla f(y+\tau(x-y)) - \nabla f(y)\|_* \cdot \|x-y\| d\tau \\ &\overset{(2),(63)}{\leq} f(y) + \langle \nabla f(y), x-y\rangle + \int_0^1 2MC_1\|x-y\| d\tau \\ &\leq f(y) + \langle \nabla f(y), x-y\rangle + 2MC_1\|x-y\|. \end{aligned}$$

8.2. Доказательство Леммы 1

Прежде всего Лемма 8 из [21] утверждает, что $F(x)$ выпуклая, дифференцируемая, а также что неравенство (15) верно. Далее используем определение $F(x)$ и теорему о среднем значении и получаем для всех $x \in X$

$$\begin{aligned} |F(x) - f(x)| &= |\mathbb{E}\left[f(x+re)\right] - f(x)| \leq \mathbb{E}\left[|f(x+re) - f(x)|\right] \\ &\leq \mathbb{E}\left[\|\nabla f(z(x, x+re))\|_2 \cdot \|re\|_2\right] \overset{(63)}{\leq} rM, \end{aligned}$$

где $z(x, x+re)$ – выпуклая комбинация x и $x+re$.

Наконец, используя симметричность распределения e и (15) имеем

$$\begin{aligned} \|\nabla F(x)\|_*^2 &= \left\|\mathbb{E}\left[\frac{n}{r} f(x+re)e\right]\right\|_*^2 \\ &= \left\|\mathbb{E}\left[\frac{n}{2r} f(x+re)e\right] + \mathbb{E}\left[\frac{n}{2r} f(x-re)\cdot(-e)\right]\right\|_*^2 \\ &= \left\|\mathbb{E}\left[\frac{n}{2r}\left(f(x+re) - f(x-re)\right)e\right]\right\|_*^2 \\ &\leq \frac{n^2}{4r^2}\mathbb{E}\left[\|\left(f(x+re) - f(x-re)\right)e\|_*^2\right] \\ &= \frac{n^2}{4r^2}\mathbb{E}\left[\left(f(x+re) - f(x-re)\right)^2 \|e\|_*^2\right] \\ &= \frac{n^2}{4r^2}\mathbb{E}\left[\left(\left(f(x+re) - \alpha\right) - \left(f(x-re) - \alpha\right)\right)^2 \|e\|_*^2\right]. \end{aligned}$$

Далее, используем (60) и получаем

$$\frac{n^2}{4r^2}\mathbb{E}\left[\left((f(x+re)-\alpha)-(f(x-re)-\alpha)\right)^2\|e\|_*^2\right]$$
$$\leq \tfrac{n^2}{2r^2}\mathbb{E}\left[\left((f(x+re)-\alpha)^2+(f(x-re)-\alpha)^2\right)\|e\|_*^2\right]$$
$$\leq \tfrac{n^2}{2r^2}\left(\mathbb{E}\left[(f(x+re)-\alpha)^2\|e\|_*^2\right]+\mathbb{E}\left[(f(x-re)-\alpha)^2\|e\|_*^2\right]\right).$$

Поскольку распределение e симметрично

$$\frac{n^2}{2r^2}\left(\mathbb{E}\left[(f(x+re)-\alpha)^2\|e\|_*^2\right]+\mathbb{E}\left[(f(x-re)-\alpha)^2\|e\|_*^2\right]\right)$$
$$= \tfrac{n^2}{r^2}\mathbb{E}\left[(f(x+re)-\alpha)^2\|e\|_*^2\right]$$
$$\overset{(62)}{\leq} \tfrac{n^2}{r^2}\sqrt{\mathbb{E}\left[\|e\|_*^4\right]}\sqrt{\mathbb{E}\left[(f(x+re)-\alpha)^4\right]}.$$

Подставляя $\alpha = \mathbb{E}[f(x+re)]$ и используя Лемму 6 вместе с фактом, что $f(x+re)$ является Mr-Липшицевой по e относительно $\|\cdot\|_2$-нормы (так как $\|\nabla f(x)\|_2 \leq M$), получаем

$$\frac{n^2p_*^2}{r^2}\sqrt{\mathbb{E}\left[(f(x+re)-\alpha)^4\right]} \leq \frac{n^2p_*^2}{r^2}\overline{c}\frac{(Mr)^2}{n} = \overline{c}np_*^2M^2,$$

где $\overline{c}$ – некоторая положительная постоянная. То есть доказано, что

$$\|\nabla F(x)\|_*^2 \leq \overline{c}np_*^2M^2,$$

что влечет (16) с $\tilde{c} = \sqrt{\overline{c}}$.

8.3. Доказательство Леммы 2

Докажем это неравенство аналогично тому, как это было сделано в Лемме 10 (смотрим [21]). Начнем с (18):

$$\begin{aligned}\mathbb{E}[\tilde{f}_r'(x)] &= \frac{n}{2r}\mathbb{E}[(\tilde{f}(x+re)-\tilde{f}(x-re))e] \\ &= \frac{n}{2r}\left(\mathbb{E}[f(x+re,\xi)e]-\mathbb{E}[f(x-re,\xi)e]+\mathbb{E}[\Delta(x+re)e]-\mathbb{E}[\Delta(x-re)e]\right)\end{aligned}$$

Принимая во внимание независимость e, ξ и (8), имеем
$\mathbb{E}[f(x+re,\xi)e] = \mathbb{E}_e\left[\mathbb{E}_\xi[f(x+re,\xi)e]\right] = \mathbb{E}_e\left[f(x+re,\xi)]\right]$. Тогда

$$\begin{aligned}\|\mathbb{E}[\tilde{f}_r'(x)]-\nabla F(x)\|_* &= \Big\|\frac{n}{2r}\big(\mathbb{E}_e\left[f(x+re)e\right]-\mathbb{E}_e\left[f(x-re)e\right] \\ &\quad +\mathbb{E}[\Delta(x+re)e]-\mathbb{E}[\Delta(x-re)e]\big)-\nabla F(x)\Big\|_* \\ &\overset{(15)}{=} \frac{n}{2r}\left\|\mathbb{E}_e[\Delta(x+re)e]-\mathbb{E}_e[\Delta(x-re)e]\right\|_* \\ &\overset{(62)}{\leq} \frac{n}{r}\sqrt{\mathbb{E}_e\left[|\Delta(x+re)|^2\right]\cdot\mathbb{E}_e\left[\|e\|_*^2\right]} \qquad (64)\end{aligned}$$

Применяя ограниченность $\Delta(x)$ и (12) к (64), мы получаем (18).

Далее докажем вторую часть леммы:

$$\begin{aligned}
\mathbb{E}[\|\tilde{f}_r'(x)\|_*^2] &= \mathbb{E}\left[\left\|\frac{n}{2r}\left(\widetilde{f}(x+re)-\widetilde{f}(x-re)\right)e\right\|_*^2\right] \\
&\overset{(60)}{\leq} \frac{n^2}{2r^2}\mathbb{E}\left[\|e\|_*^2\left(f(x+re,\xi)-f(x-re,\xi)\right)^2\right] \\
&\quad +\frac{n^2}{2r^2}\mathbb{E}\left[\|e\|_*^2\left(\Delta(x+re)-\Delta(x-re)\right)^2\right] \\
&\overset{(60)}{\leq} \frac{n^2}{2r^2}\mathbb{E}\left[\|e\|_*^2\left((f(x+re,\xi)-\alpha)-(f(x-re,\xi)-\alpha)\right)^2\right] \\
&\quad +\frac{n^2}{r^2}\mathbb{E}\left[\|e\|_*^2\left(\Delta^2(x+re)+\Delta^2(x-re)\right)\right] \\
&\overset{(7),(62)}{\leq} \frac{n^2}{r^2}\mathbb{E}_e\left[\|e\|_*^2\left(\left(f(x+re,\xi)-\alpha\right)^2+\left(f(x-re,\xi)-\alpha\right)^2\right)\right] \\
&\quad +\frac{2n^2\Delta^2}{r^2}\sqrt{\mathbb{E}\left[\|e\|_*^4\right]} \\
&\overset{(12)}{\leq} \frac{n^2}{r^2}\left(\mathbb{E}\left[\|e\|_*^2\left(f(x+re,\xi)-\alpha\right)^2\right]+\mathbb{E}\left[\|e\|_*^2\left(f(x-re,\xi)-\alpha\right)^2\right]\right) \\
&\quad +\frac{2n^2p_*^2\Delta^2}{r^2}. \qquad (65)
\end{aligned}$$

Так как распределение e симметрично, можно переписать (65) в следующем виде:

$$\begin{aligned}
&\frac{n^2}{r^2}\left(\mathbb{E}\left[\|e\|_*^2\left(f(x+re,\xi)-\alpha\right)^2\right]+\mathbb{E}\left[\|e\|_*^2\left(f(x-re,\xi)-\alpha\right)^2\right]\right)+\frac{2n^2p_*^2\Delta^2}{r^2} \\
&\qquad = \tfrac{2n^2}{r^2}\mathbb{E}\left[\|e\|_*^2\left(f(x+re,\xi)-\alpha\right)^2\right]+\tfrac{2n^2p_*^2\Delta^2}{r^2}.
\end{aligned}$$

Используя независимость e и ξ, получаем

$$\begin{aligned}
&\frac{2n^2}{r^2}\mathbb{E}\left[\|e\|_*^2\left(f(x+re,\xi)-\alpha\right)^2\right]+\frac{2n^2p_*^2\Delta^2}{r^2} \\
&\qquad = \tfrac{2n^2}{r^2}\mathbb{E}_\xi\left[\mathbb{E}_e\left[\|e\|_*^2\left(f(x+re,\xi)-\alpha\right)^2\right]\right]+\tfrac{2n^2p_*^2\Delta^2}{r^2}.
\end{aligned}$$

Применяя неравенство Коши-Буняковского-Шварца, (12) имеем

$$\begin{aligned}
&\frac{2n^2}{r^2}\mathbb{E}_\xi\left[\mathbb{E}_e\left[\|e\|_*^2\left(f(x+re,\xi)-\alpha\right)^2\right]\right]+\frac{2n^2p_*^2\Delta^2}{r^2} \\
&\qquad \overset{(62)}{\leq} \tfrac{2n^2}{r^2}\mathbb{E}_\xi\left[\sqrt{\mathbb{E}_e\left[\|e\|_*^4\right]}\sqrt{\mathbb{E}_e\left[\left(f(x+re,\xi)-\alpha\right)^4\right]}\right]+\tfrac{2n^2p_*^2\Delta^2}{r^2} \\
&\qquad \leq \tfrac{2n^2p_*^2}{r^2}\mathbb{E}_\xi\left[\sqrt{\mathbb{E}_e\left[\left(f(x+re,\xi)-\alpha\right)^4\right]}\right]+\tfrac{2n^2p_*^2\Delta^2}{r^2}.
\end{aligned}$$

Подставляя $\alpha=\mathbb{E}_e[f(x+re,\xi)]$ и используя Лемму 6 вместе с $rM(\xi)$-Липшецовостью функции $f(x+re,\xi)$ по e относительно $\|\cdot\|_2$-нормы, можно доказать

$$\begin{aligned}
\frac{2n^2p_*^2}{r^2}\mathbb{E}_\xi\left[\sqrt{\mathbb{E}_e\left[\left(f(x+re,\xi)-\alpha\right)^4\right]}\right]+\frac{2n^2p_*^2\Delta^2}{r^2} &\leq \frac{2n^2p_*^2}{r^2}\mathbb{E}_\xi\left[c\frac{r^2M^2(\xi)}{n}\right]+\frac{2n^2p_*^2\Delta^2}{r^2} \\
&\overset{(9)}{=} 2p_*^2\left(cnM^2+\frac{n^2\Delta^2}{r^2}\right),
\end{aligned}$$

где c–некоторая положительная константа.

8.4. Доказательство Леммы 3

Доказательство этой леммы полностью повторяет доказательство Предложения 8.3 из [2]. Тем не менее, докажем его. Рассмотрим следующие функции:

$$l_F(u_{t-1}, u) = F(u_{t-1}) + \langle \nabla F_r(u_{t-1}), u - u_{t-1} \rangle,$$

$$\tilde{l}_F(u_{t-1}, u) = F(u_{t-1}) + \langle \widetilde{f}'_r(u_{t-1}), u - u_{t-1} \rangle.$$

Эти определения подразумевают, что $\tilde{l}_F(u_{t-1}, u) - l_F(u_{t-1}, u) = \langle \delta_t, u - u_{t-1} \rangle$, где δ_t определена в (26). Леммы 9 и 1 дадут $F(u_t) \leq l_F(u_{t-1}, u_t) + \tilde{M}\|u_t - u_{t-1}\|$, где $\tilde{M} = c\sqrt{n}C_1 M$. Применяя дополнительно $h(u_t)+\beta V(x, u_t)$ к этому неравенству и используя (25), имеем

$$\Phi(u_t) \leq h(u_t) + l_F(u_{t-1}, u_t) + \beta V(x, u_t) + \tilde{M}\|u_t - u_{t-1}\|.$$

Из $\tilde{l}_F(u_{t-1}, u) - l_F(u_{t-1}, u) = \langle \delta_t, u - u_{t-1} \rangle$ получаем

$$\begin{aligned} \Phi(u_t) &\leq h(u_t) + l_F(u_{t-1}, u_t) + \beta V(x, u_t) + \tilde{M}\|u_t - u_{t-1}\| \\ &= h(u_t) + \tilde{l}_F(u_{t-1}, u_t) - \langle \delta_t, u_t - u_{t-1} \rangle + \beta V(x, u_t) + \tilde{M}\|u_t - u_{t-1}\| \\ &\overset{(61)}{\leq} h(u_t) + \tilde{l}_F(u_{t-1}, u_t) + \beta V(x, u_t) + (\tilde{M} + \|\delta_t\|_*)\|u_t - u_{t-1}\|. \end{aligned}$$

Применяя Лемму 7 к (20), замечаем, что для любых $u \in X$

$$\begin{aligned} h(u_t) + \tilde{l}_F(u_{t-1}, u_t) + \beta V(x, u_t) + \beta p_t V(u_{t-1}, u_t) \\ \leq h(u) + \tilde{l}_F(u_{t-1}, u) + \beta V(x, u) + \beta p_t V(u_{t-1}, u) - \beta(1 + p_t)V(u_t, u) \\ = h(u) + l_F(u_{t-1}, u) + \langle \delta_t, u - u_{t-1} \rangle + \beta V(x, u) \\ + \beta p_t V(u_{t-1}, u) - \beta(1 + p_t)V(u_t, u) \\ \leq \Phi(u) + \beta p_t V(u_{t-1}, u) - \beta(1 + p_t)V(u_t, u) + \langle \delta_t, u - u_{t-1} \rangle, \end{aligned}$$

где последнее неравенство следует из выпуклости F (смотри Лемму 1) и (25). Кроме того, сильная выпуклость V влечет

$$\begin{aligned} -\beta p_t V(u_{t-1}, u_t) + (\tilde{M} + \|\delta_t\|_*)\|u_t - u_{t-1}\| &\leq -\frac{\beta p_t}{2}\|u_t - u_{t-1}\|^2 + (\tilde{M} + \|\delta_t\|_*)\|u_t - u_{t-1}\| \\ &\leq \frac{\left(\tilde{M} + \|\delta_t\|_*\right)^2}{2\beta p_t}, \end{aligned}$$

где последнее неравенство следует из простого факта, что $-{at^2}/{2} + bt \leq {b^2}/{(2a)}$ для любого $a > 0$.

Объединяя предыдущие три неравенства, заключаем, что

$$\Phi(u_t) - \Phi(u) \leq \beta p_t V(u_{t-1}, u) - \beta(1+p_t)V(u_t, u) + \frac{\left(\tilde{M} + \|\delta_t\|_*\right)^2}{2\beta p_t} + \langle \delta_t, u - u_{t-1}\rangle.$$

Теперь, разделив обе части вышеприведенного неравенства на $1 + p_t$ и переставив члены, получим

$$\beta V(u_t, u) + \frac{\Phi(u_t) - \Phi(u)}{1+p_t} \leq \frac{\beta p_t}{1+p_t} V(u_{t-1}, u) + \frac{\left(\tilde{M} + \|\delta_t\|_*\right)^2}{2\beta(1+p_t)p_t} + \frac{\langle \delta_t, u - u_{t-1}\rangle}{1+p_t},$$

что вместе с Леммой 8 дает

$$\frac{\beta}{P_t} V(u_t, u) + \sum_{i=1}^{t} \frac{\Phi(u_i) - \Phi(u)}{P_i(1+p_i)} \leq \beta V(u_0, u) + \sum_{i=1}^{t} \left[\frac{\left(\tilde{M} + \|\delta_i\|_*\right)^2}{2\beta P_i(1+p_i)p_i} + \frac{\langle \delta_i, u - u_{i-1}\rangle}{P_i(1+p_i)} \right]. \quad (66)$$

По определению $\tilde{u}_t$ (смотри (21)) и (23) имеем

$$\begin{aligned} \tilde{u}_t &= \frac{P_t}{1-P_t}\left(\frac{1-P_{t-1}}{P_{t-1}}\tilde{u}_{t-1} + \frac{1}{P_t(1+p_t)}u_t\right), \\ \tilde{u}_t &= \frac{P_t}{1-P_t}\left(\frac{1-P_{t-2}}{P_{t-2}}\tilde{u}_{t-2} + \frac{1}{P_{t-1}(1+p_{t-1})}u_{t-1} + \frac{1}{P_t(1+p_t)}u_t\right) \\ &= \ldots = \frac{P_t}{1-P_t}\sum_{i=1}^{t}\frac{1}{P_i(1+p_i)}u_i. \end{aligned} \quad (67)$$

Комбинируя (66) и (67), завершаем доказательство.

8.5. Доказательство Теоремы 1

Доказательство этой теоремы практически идентично доказательству теоремы 8.2 из [2], и, выполнив аналогичные шаги, можно получить итоговое неравенство, которое является аналогом неравенства (8.1.69) из [2]. Для удобства ниже приводим полное доказательство.

Используя (24), определение Φ_k и $(x_k, \tilde{x}_k)$, имеем для любого $u \in X$

$$\begin{aligned} &\beta_k(1-P_{T_k})^{-1}V(x_k, u) + [\Phi_k(\tilde{x}_k) - \Phi_k(u)] \\ &\qquad \leq \beta_k P_{T_k}(1-P_{T_k})^{-1}V(x_{k-1}, u) + \frac{P_{T_k}}{1-P_{T_k}}\sum_{i=1}^{T_k}\frac{\frac{(\tilde{M}+\|\delta_{k,i}\|_*)^2}{2\beta_k p_i} + \langle \delta_{k,i}, u - u_{k,i-1}\rangle}{p_i P_{i-1}} \end{aligned} \quad (68)$$

Во-первых, обратим внимание, что по определению $\overline{x}_k$ и $\underline{x}_k$, получаем $\overline{x}_k - \underline{x}_k = \gamma_k(\tilde{x}_k - x_{k-1})$. Используя это наблюдение, L-гладкость g (смотри (3)), определение l_g из (22) и выпуклость g, имеем

$$\begin{aligned} g(\overline{x}_k) &\leq l_g(\underline{x}_k, \overline{x}_k) + \frac{L}{2}\|\overline{x}_k - \underline{x}_k\|^2 \\ &= (1-\gamma_k) l_g(\underline{x}_k, \overline{x}_{k-1}) + \gamma_k l_g(\underline{x}_k, \tilde{x}_k) + \frac{L\gamma_k^2}{2}\|\tilde{x}_k - x_{k-1}\|^2 \\ &\leq (1-\gamma_k) g(\overline{x}_{k-1}) + \gamma_k \left[l_g(\underline{x}_k, \tilde{x}_k) + \beta_k V(x_{k-1}, \tilde{x}_k)\right] \\ &\quad -\gamma_k\beta_k V(x_{k-1}, \tilde{x}_k) + \frac{L\gamma_k^2}{2}\|\tilde{x}_k - x_{k-1}\|^2 \\ &\leq (1-\gamma_k) g(\overline{x}_{k-1}) + \gamma_k \left[l_g(\underline{x}_k, \tilde{x}_k) + \beta_k V(x_{k-1}, \tilde{x}_k)\right] \\ &\quad - \left(\gamma_k\beta_k - L\gamma_k^2\right) V(x_{k-1}, \tilde{x}_k) \\ &\leq (1-\gamma_k) g(\overline{x}_{k-1}) + \gamma_k \left[l_g(\underline{x}_k, \tilde{x}_k) + \beta_k V(x_{k-1}, \tilde{x}_k)\right], \end{aligned}$$

где третье неравенство следует из сильной выпуклости V, а последнее неравенство следует из (30). По выпуклости F имеем

$$F(\overline{x}_k) \leq (1-\gamma_k) F(\overline{x}_{k-1}) + \gamma_k F(\tilde{x}_k).$$

Суммируя предыдущие два неравенства и используя определения Ψ и $\Phi_k(\tilde{x}_k) = F(\tilde{x}_k) + l_g(\underline{x}_k, \tilde{x}_k) + \beta_k V(x_{k-1}, \tilde{x}_k)$, получаем

$$\Psi(\overline{x}_k) \leq (1-\gamma_k)\Psi(\overline{x}_{k-1}) + \gamma_k \Phi_k(\tilde{x}_k).$$

Вычитая $\Psi(u)$ из обеих частей вышеприведенного неравенства, имеем

$$\Psi(\overline{x}_k) - \Psi(u) \leq (1-\gamma_k)[\Psi(\overline{x}_{k-1}) - \Psi(u)] + \gamma_k[\Phi_k(\tilde{x}_k) - \Psi(u)].$$

Также отметим, что по определению Φ_k и выпуклости g,

$$\Phi_k(u) \leq F(u) + g(u) + \beta_k V(x_{k-1}, u) = \Psi(u) + \beta_k V(x_{k-1}, u), \quad \forall u \in X.$$

Объединяя эти два неравенства, получаем для всех $u \in X$

$$\Psi(\overline{x}_k) - \Psi(u) \leq (1-\gamma_k)[\Psi(\overline{x}_{k-1}) - \Psi(u)] + \gamma_k[\Phi_k(\tilde{x}_k) - \Phi_k(u) + \beta_k V(x_{k-1}, u)]. \quad (69)$$

Используя (68) и (69), имеем для всех $u \in X$

$$\begin{aligned} \Psi(\overline{x}_k) - \Psi(u) &\leq (1-\gamma_k)[\Psi(\overline{x}_{k-1}) - \Psi(u)] + \gamma_k \left\{ \frac{\beta_k}{1-P_{T_k}}[V(x_{k-1}, u) - V(x_k, u)] \right. \\ &\quad \left. + \frac{P_{T_k}}{1-P_{T_k}} \sum_{i=1}^{T_k} \frac{1}{p_i P_{i-1}} \left[\frac{\left(\tilde{M} + \|\delta_{k,i}\|_*\right)^2}{2\beta_k p_i} + \langle \delta_{k,i}, u - u_{k,i-1} \rangle \right] \right\}. \quad (70) \end{aligned}$$

По приведенному выше неравенство и Лемме 8, заключаем, что для любого $u \in X$

$$\begin{aligned}\Psi(\overline{x}_N) - \Psi(u) &\leq \Gamma_N(1-\gamma_1)[\Psi(\overline{x}_0) - \Psi(u)] \\ &+\Gamma_N \sum_{k=1}^{N} \frac{\beta_k\gamma_k}{\Gamma_k(1-P_{T_k})}\left[V(x_{k-1},u) - V(x_k,u)\right] \\ &+\Gamma_N \sum_{k=1}^{N} \frac{\gamma_k P_{T_k}}{\Gamma_k(1-P_{T_k})}\cdot \\ &\cdot \sum_{i=1}^{T_k} \frac{1}{p_i P_{i-1}}\left[\frac{\left(\tilde{M}+\|\delta_{k,i}\|_*\right)^2}{2\nu\beta_k p_i} + \langle \delta_{k,i}, u - u_{k,i-1}\rangle\right]. \end{aligned} \tag{71}$$

Из (28) следует, что для всех $u \in X$

$$\begin{aligned}\sum_{k=1}^{N} \frac{\beta_k\gamma_k}{\Gamma_k(1-P_{T_k})}\left[V(x_{k-1},u) - V(x_k,u)\right] &\leq \frac{\beta_1\gamma_1}{\Gamma_1(1-P_{T_1})}V(x_0,u) - \frac{\beta_N\gamma_N}{\Gamma_N(1-P_{T_N})}V(x_N,u) \\ &\leq \frac{\beta_1}{1-P_{T_1}}V(x_0,u), \end{aligned} \tag{72}$$

где последнее неравенство следует из того факта, что $\gamma_1 = \Gamma_1 = 1$, $P_{T_N} \leq 1$, и $V(x_N,u) \geq 0$. Неравенство (72) и тот факт, что $\gamma_1 = 1$ вместе с неравенством (71) влекут, что для всех $u \in X$

$$\begin{aligned}\Psi(\overline{x}_N) - \Psi(u) &\leq \frac{\beta_k}{1-P_{T_1}}V(x_0,u) \\ &+\Gamma_N \sum_{k=1}^{N} \frac{\gamma_k P_{T_k}}{\Gamma_k(1-P_{T_k})}\cdot \\ &\cdot \sum_{i=1}^{T_k} \frac{1}{p_i P_{i-1}}\left[\frac{(\tilde{M}^2+\|\delta_{k,i}\|_*^2)}{\beta_k p_i} + \langle \delta_{k,i}, u - u_{k,i-1}\rangle\right]. \end{aligned} \tag{73}$$

Далее покажем, что

$$\mathbb{E}[||\delta_{k,i}||_*^2] \leq \sigma^2 \tag{74}$$

для σ^2 из (31). Для любого $x \in X$ имеем

$$\begin{aligned}\mathbb{E}[\|\delta\|_*^2] &= \mathbb{E}[\|\tilde{f}_r'(x) - \nabla F(x)\|_*^2] \leq 2\mathbb{E}\|\tilde{f}_r'(x)\|_*^2 + 2\mathbb{E}\|\nabla F(x)\|_*^2 \\ &\overset{(19)}{\leq} 4p_*^2\left(cnM^2 + \frac{n^2\Delta^2}{r^2}\right) + 2\|\nabla F(x)\|_*^2 \\ &\overset{(16)}{=} 4p_*^2\left(cnM^2 + \frac{n^2\Delta^2}{r^2}\right) + 2\tilde{c}^2 np_*^2M^2 \\ &= 4p_*^2\left(CnM^2 + \frac{n^2\Delta^2}{r^2}\right), \end{aligned}$$

где $C = c + \tilde{c}^2/2$. Для скалярного произведения имеем следующую оценку:

$$\begin{aligned}\mathbb{E}[\langle \delta_{k,i}, u - u_{k,i-1}\rangle] &\overset{(6)}{=} \frac{n}{2r}\mathbb{E}[\langle \Delta_{k,i} e_{k,i}, u - u_{k,i-1}\rangle] \\ &\leq \frac{n}{2r}\mathbb{E}[|\Delta_{k,i}| \cdot \|e_{k,i}\|_* \cdot \|u - u_{k,i-1}\|] \overset{(7),(12)}{\leq} \frac{\Delta n D_X p_*}{r}.\end{aligned}$$

Беря математическое ожидание с обеих сторон (73) и используя (74) и (75), получаем (29).

8.6. Доказательство Следствия 1

Используя рекурренту (23) и (32), замечаем, что

$$P_t = \frac{2}{(t+1)(t+2)}, \tag{75}$$

$$P_{T_k} \leq P_{T_{k-1}} \leq \ldots \leq P_{T_1} \leq \frac{1}{3} \tag{76}$$

и из соотношений (30) и (33) получаем, что

$$\Gamma_k = \frac{2}{k(k+1)}, \tag{77}$$

откуда следует (27).

Из (33), (75), (76) выводим (28). Простые вычисления и соотношения (32), (75) дают

$$\sum_{i=1}^{T_k} \frac{1}{p_i^2 P_{i-1}} = 2\sum_{i=1}^{T_k} \frac{i+1}{i} \leq 4T_k, \tag{78}$$

$$\sum_{i=1}^{T_k} \frac{1}{p_i P_{i-1}} = \frac{1}{2}\sum_{i=1}^{T_k} i = \frac{1}{4}T_k(T_k+1). \tag{79}$$

Далее из (33), (77), (78), (79) можно вывести, что

$$\sum_{i=1}^{T_k} \frac{\gamma_k P_{T_k}}{\Gamma_k \beta_k (1-P_{T_k}) p_i^2 P_{i-1}} \leq \frac{4\gamma_k P_{T_k} T_k}{\Gamma_k \beta_k (1-P_{T_k})} = \frac{4k^2}{L(T_k+3)}, \tag{80}$$

$$\sum_{i=1}^{T_k} \frac{\gamma_k P_{T_k}}{\Gamma_k (1-P_{T_k}) p_i P_{i-1}} \leq \frac{\gamma_k P_{T_k} T_k (T_k+1)}{4\Gamma_k (1-P_{T_k})} = \frac{(T_k+1)k}{2(T_k+3)}. \tag{81}$$

Наконец, неравенства (29), (76), (77), (80), (81) влекут

$$\begin{aligned}\mathbb{E}[\Psi(\overline{x}_N) - \Psi(x^*)] &\leq \frac{2L}{N(N+1)}[3V(x_0,x^*) + 4\tilde{D}] + \frac{2n\Delta D_X p_*}{rN(N+1)}\sum_{k=1}^{N}\frac{(T_k+1)k}{(T_k+3)} \\ &\leq \frac{2L}{N(N+1)}[3V(x_0,x^*) + 4\tilde{D}] + \frac{2n\Delta D_X p_*}{rN(N+1)}\sum_{k=1}^{N} k \\ &= \frac{2L}{N(N+1)}[3V(x_0,x^*) + 4\tilde{D}] + \frac{n\Delta D_X p_*}{r}.\end{aligned} \tag{82}$$

8.7. Доказательство Следствия 2

Доказательство (35) следует из (34) и (14). Используя (35) и (36), получаем (37). Наконец,

$$\begin{aligned}\sum_{i=1}^{N} T_k &\overset{(33)}{\leq} \sum_{i=1}^{N}\left(\frac{4N(\tilde{M}^2+\sigma^2)k^2}{3D_{X,V}^2L^2}+1\right)\\ &= \frac{2}{18}\frac{N^2(N+1)(2N+1)(\tilde{M}^2+\sigma^2)}{D_{X,V}^2L^2}+N\\ &\leq \frac{2}{9}\frac{(N+1)^4(\tilde{M}^2+\sigma^2)}{D_{X,V}^2L^2}+N \qquad (83)\end{aligned}$$

и

$$\begin{aligned}\sigma^2 &= O\left(p_*^2\left(CnM^2+\frac{n^2\Delta^2}{r^2}\right)\right)\\ &= O\left(p_*^2nM^2\right), \qquad (84)\end{aligned}$$

$$\tilde{M}^2 = O\left(nC_1^2M^2\right), \qquad (85)$$

где мы использовали $\varepsilon = O\left(\sqrt{n}MD_X\right)$. Из (37) известно, что $N = O\left(\sqrt{LD_{X,V}^2/\varepsilon}\right)$, вместе с (83), (84), (85) это дает (38).

9. Пропущенные доказательства из Раздела 2.3

9.1. Доказательство Теоремы 2

Докажем этот результат по индукции. Из (82) имеем

$$\mathbb{E}[\Psi(y_i)-\Psi(y^*)\mid y_{i-1}] \leq \frac{2L}{N_0(N_0+1)}\left(3V(y_{i-1},y^*)+4\tilde{D}\right)+\frac{n\Delta D_X p_*}{r}.$$

Здесь использовалось, что y_{i-1} равно x_0, а y_i – выход для `M-zoSA` после i-ой итерации. Поскольку Ψ является суммой выпуклой и μ-сильно выпуклой функции, имеем, что Ψ является μ-сильно выпуклой и

$$\mathbb{E}[\Psi(y_i)-\Psi(y^*)\mid y_{i-1}] \leq \frac{2L}{N_0(N_0+1)}\left(\frac{3}{\mu}\left(\Psi(y_{i-1})-\Psi(y^*)\right)+4\tilde{D}\right)+\frac{n\Delta D_X p_*}{r}.$$

Взяв полное математическое ожидание от обеих сторон предыдущего неравенства и используя гипотезу индукции и определение $\tilde{D}$, получаем, что

$$\mathbb{E}[\Psi(y_i)-\Psi(y^*)] \leq \frac{2L}{N_0^2}\frac{5\rho_0}{\mu 2^{i-1}}+\frac{2L}{N_0^2}\frac{6n\Delta D_X}{\mu r}+\frac{n\Delta D_X}{r} \leq \frac{\rho_0}{2^i}+\frac{2n\Delta D_X p_*}{r},$$

где последнее неравенство следует из определения N_0.

9.2. Доказательство Следствия 3

Объединяя (14) и (39), имеем (40). Далее если r и Δ удовлетворяют (41), то $2rM + \frac{2n\Delta D_X p_*}{r} = O(\mathbf{e})$. Рассчитывая общее число рестартов, $I = \lceil \log_2 \max[1, \rho_0/\varepsilon] \rceil$ получаем $\frac{\rho_0}{2^I} = O(\varepsilon)$, и, как следствие, $\mathbb{E}[\Psi_0(y_i) - \Psi_0(y^*)] = O(\varepsilon)$. Поэтому общее число вычислений ∇g, равное $N_0 I$ и ограничено (42).

Теперь вычислим общее количество вычислений $\tilde{f}'_r$. Без ограничения общности предположим, что $\rho_0 > \varepsilon$. Используя предыдущую оценку I и определение T_k, получим

$$\begin{aligned}
\sum_{i=1}^{I}\sum_{k=1}^{N_0} T_k &\leq \sum_{i=1}^{I}\sum_{k=1}^{N_0}\left(\frac{\mu N_0(\tilde{M}^2+\sigma^2)k^2}{\rho_0 L^2}2^i + 1\right) \\
&\leq \sum_{i=1}^{I}\left[\frac{\mu N_0(\tilde{M}^2+\sigma^2)}{3\rho_0 L^2}(N_0+1)^3 2^i + N_0\right] \\
&\leq \frac{\mu N_0(N_0+1)^3(\tilde{M}^2+\sigma^2)}{3\rho_0 L^2}2^{I+1} + N_0 I \\
&\leq \frac{4\mu(N_0+1)^4(\tilde{M}^2+\sigma^2)}{3\varepsilon L^2} + N_0 I.
\end{aligned}$$

Это неравенство, определение N_0, неравенства (84) и (85) дают оценку (43) для общего числа вычислений $\tilde{f}'_r$.

10. Пропущенные доказательства из Раздела 4

10.1. Доказательство Леммы 4

Пусть x^* – решение исходной задачи, а $\hat{x}$ – точка множества X_α, ближайшая к x^*. Используем, что $f(\tilde{x}) \leq f(\hat{x})$, и M-Липшецевость функции f:

$$\begin{aligned}
f(x_k) - f(x^*) &= f(x_k) - f(\hat{x}) + f(\hat{x}) - f(x^*) \\
&\leq f(x_k) - f(\tilde{x}) + f(\hat{x}) - f(x^*) \leq \frac{\varepsilon}{2} + M\|\hat{x} - x^*\|_2 \\
&\leq \frac{\varepsilon}{2} + Mr. \qquad (86)
\end{aligned}$$

10.2. Доказательство Леммы 5

Без ограничения общности можно перенести центр сфер в ноль, а также, в силу симметрии, рассмотреть только те части сфер, где все компоненты положительны.

Тогда переписав задачу нахождения минимального расстояния, получим

$$\begin{aligned} \min_{x,y\in\mathbb{R}^n_+} \quad & \|x-y\|_2 \\ \text{s.t.} \quad & x \in \mathcal{S}^n_p(R,0), \\ & y \in \mathcal{S}^n_p(R(1-\alpha),0). \end{aligned} \tag{87}$$

Функция Лагранжа (87):

$$L = \sum_{i=1}^n (x_i - y_i)^2 + \lambda_1 \left(\sum_{i=1}^n x_i^p - R^p \right) + \lambda_2 \left(\sum_{i=1}^n y_i^p - (1-\alpha)^p R^p \right).$$

Обратим внимание, что мы не добавляем ограничения для $x_i \geq 0$ и $y_i \geq 0$ в функцию Лагранжа.

Взяв производные по x_i и y_i и используя необходимые условия экстремума:

$$\begin{cases} L_{x_i} = 2(x_i - y_i) + \lambda_1 p x_i^{p-1} = 0, \\ L_{y_i} = 2(y_i - x_i) + \lambda_2 p y_i^{p-1} = 0, \\ \sum\limits_{i=1}^n x_i^p - R^p = 0, \\ \sum\limits_{i=1}^n y_i^p - (1-\alpha)^p R^p = 0. \end{cases} \tag{88}$$

Можно заметить, что $x_i > y_i$, следовательно $\lambda_1 < 0$, $\lambda_2 > 0$. Из первых двух уравнений (88):

$$-\lambda_1 p x_i^{p-1} = \lambda_2 p y_i^{p-1} \tag{89}$$

$$\begin{aligned} (-\lambda_1)^{p/p-1} x_i^p &= \lambda_2^{p/p-1} y_i^p \\ (-\lambda_1)^{p/p-1} \sum_{i=1}^n x_i^p &= \lambda_2^{p/p-1} \sum_{i=1}^n y_i^p \\ (-\lambda_1)^{p/p-1} R^p &= \lambda_2^{p/p-1} (1-\alpha)^p R^p \\ -\lambda_1 &= \lambda_2 (1-\alpha)^{p-1}. \end{aligned} \tag{90}$$

Объединяя (89) и (90), получаем

$$(1-\alpha) x_i = y_i. \tag{91}$$

Подставляя y_i из первого уравнения (88) во второе уравнение (88) и используя (89),

имеем

$$\begin{aligned}
-\lambda_1 x_i^{p-1} &= \lambda_2 \left(x_i + \frac{\lambda_1 p}{2} x_i^{p-1} \right)^{p-1} \\
(1-\alpha)^{p-1} x_i^{p-1} &= \left(x_i + \frac{\lambda_1 p}{2} x_i^{p-1} \right)^{p-1} \\
(1-\alpha) x_i &= x_i + \frac{\lambda_1 p}{2} x_i^{p-1} \\
-\alpha &= \frac{\lambda_1 p}{2} x_i^{p-2} \\
\left(\frac{-2\alpha}{\lambda_1 p} \right)^{p/p-2} &= x_i^p \\
n \left(\frac{-2\alpha}{\lambda_1 p} \right)^{p/p-2} &= R^p \\
\lambda_1 &= \frac{-2\alpha n^{p-2/p}}{p R^{p-2}}.
\end{aligned}$$

Тогда по (90):

$$\lambda_2 = \frac{2\alpha n^{p-2/p}}{(1-\alpha)^{p-1} p R^{p-2}}.$$

Подставляя λ_1, $x_i - y_i = \alpha x_i$ в первое уравнение (88):

$$\alpha x_i = \frac{2\alpha n^{p-2/p} p}{2 p R^{p-2}} x_i^{p-1}.$$

Откуда следует, что

$$x_i = \frac{R}{n^{1/p}}, \quad y_i = \frac{(1-\alpha)R}{n^{1/p}}.$$

Найденные значения неотрицательны. Легко получить значение m:

$$m = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} = \frac{\alpha R}{n^{1/p - 1/2}}.$$

Осталось только убедиться, что найденное значение является минимальным. Учитывая, что $x_i = x_j$, $y_i = y_j$ и $y_i = (1-\alpha)x_i$, можно записать $dx_i = dx_j$, $y_i = y_j$, $dy_i = (1-\alpha)dx_i$ и найти d^2L:

$$\begin{aligned}
d^2 L &= \sum_{i=1}^{n} L_{x_i x_i} (dx_i)^2 + 2 \sum_{i=1}^{n} L_{x_i y_i} dx_i dy_i + \sum_{i=1}^{n} L_{y_i y_i} (dy_i)^2 \\
&= n L_{x_1 x_1} (dx_1)^2 + 2n L_{x_1 y_1} dx_i dy_i + n L_{y_1 y_1} (dy_1)^2 \\
&= n (L_{x_1 x_1} + 2(1-\alpha) L_{x_1 y_1} + (1-\alpha)^2 L_{y_1 y_1}) (dx_1)^2 \\
&= n \left(2 - 2\alpha(p-1) - 4(1-\alpha) + (1-\alpha)(2 + 2\alpha(p-1))\right) (dx_1)^2 \\
&= 2n\alpha(1 - \alpha(p-1))(dx_1)^2.
\end{aligned}$$

Для $\alpha \in (0;1)$ и $p \in [1;2]$ имеем $\alpha(p-1) \leq 0$, следовательно, $d^2L \geq 0$. Это означает, что найден именно минимум.

11. Заключение

В заключение, предложенный метод — `zoSA` — является первым, насколько известно, $1/2$-методом для выпуклой композитной оптимизации: он использует оракул нулевого порядка для негладкого члена, и оракул первого порядка – для гладкого. Метод имеет хорошо изученную теорию и конкурентоспособен на практике даже с некоторыми методами первого порядка.

Что касается развития идей данной части работы, было бы интересно изучить распределенные методы нулевого порядка для гладкой децентрализованной распределенной оптимизации, используя идеи из [12]. Другое направление для будущих исследований заключается в разработке анализа предложенного метода для случая, когда множество X не ограничено, и, в частности, когда $X = \mathbb{R}^n$. Это можно сделать с помощью метода рекуррент из [12, 28].

Также была исследована работа безградиентых методов в случае, когда нельзя выходить за допустимое множестве: предложен общий алгоритм действий, уточнены оценки на параметры аппроксимации градиента и уровень шума. Что тоже встречается впервые в литературе.

Хотелось бы рассмотреть случаи, когда допустимое множество представляет собой произведение, пересечение, объединение двух множеств, а также обобщить эту в теорию, которая будет применима для любого метода, использующего оракул нулевого порядка.

Список литературы

1. Lan G. Gradient sliding for composite optimization // Mathematical Programming. 2016. Vol. 159, no. 1-2. P. 201–235.
2. Lan G. Lectures on Optimization Methods for Machine Learning. H. Milton Stewart School of Industrial and Systems Engineering Georgia Institute of Technology, Atlanta, GA, 2019.
3. Shalev-Shwartz S., Ben-David S. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.
4. Spokoiny V. et al. Parametric estimation. Finite sample theory // The Annals of Statistics. 2012. Vol. 40, no. 6. P. 2877–2909.
5. Rao A. V. A survey of numerical methods for optimal control // Advances in the Astronautical Sciences. 2009. Vol. 135, no. 1. P. 497–528.
6. Lan G., Lee S., Zhou Y. Communication-efficient algorithms for decentralized and stochastic optimization // Mathematical Programming. 2017. P. 1–48.
7. Scaman K., Bach F., Bubeck S. et al. Optimal algorithms for smooth and strongly convex distributed optimization in networks // Proceedings of the 34th International Conference on Machine Learning-Volume 70 / JMLR. org. 2017. P. 3027–3036.
8. Scaman K., Bach F., Bubeck S. et al. Optimal algorithms for non-smooth distributed optimization in networks // Advances in Neural Information Processing Systems. 2018. P. 2745–2754.
9. Scaman K., Bach F., Bubeck S. et al. Optimal Convergence Rates for Convex Distributed Optimization in Networks // Journal of Machine Learning Research. 2019. Vol. 20, no. 159. P. 1–31.
10. Dvinskikh D., Gorbunov E., Gasnikov A. et al. On Primal and Dual Approaches for Distributed Stochastic Convex Optimization over Networks // 2019 IEEE 58th Conference on Decision and Control (CDC). 2019. P. 7435–7440.
11. Dvinskikh D., Gasnikov A. Decentralized and Parallelized Primal and Dual Accelerated Methods for Stochastic Convex Programming Problems // arXiv preprint arXiv:1904.09015. 2019.
12. Gorbunov E., Dvinskikh D., Gasnikov A. Optimal Decentralized Distributed Algorithms for Stochastic Convex Optimization // arXiv preprint arXiv:1911.07363. 2019.

13. Uribe C. A., Lee S., Gasnikov A., Nedić A. A dual approach for optimal algorithms in distributed optimization over networks // Optimization Methods and Software. 2020. Vol. 0, no. 0. P. 1–40. https://doi.org/10.1080/10556788.2020.1750013. URL: https://doi.org/10.1080/10556788.2020.1750013.
14. Stich S. U. Local SGD converges fast and communicates little // arXiv preprint arXiv:1805.09767. 2018.
15. Karimireddy S. P., Rebjock Q., Stich S. U., Jaggi M. Error feedback fixes signsgd and other gradient compression schemes // arXiv preprint arXiv:1901.09847. 2019.
16. Alistarh D., Grubic D., Li J. et al. QSGD: Communication-efficient SGD via gradient quantization and encoding // Advances in Neural Information Processing Systems. 2017. P. 1709–1720.
17. Wen W., Xu C., Yan F. et al. Terngrad: Ternary gradients to reduce communication in distributed deep learning // Advances in Neural Information Processing Systems. 2017. P. 1509–1519.
18. Bertsekas D. P., Tsitsiklis J. N. Parallel and distributed computation: numerical methods. Prentice hall Englewood Cliffs, NJ, 1989. Vol. 23.
19. Rogozin A., Gasnikov A. Projected Gradient Method for Decentralized Optimization over Time-Varying Networks // arXiv preprint arXiv:1911.08527. 2019.
20. Nesterov Y., Spokoiny V. G. Random Gradient-Free Minimization of Convex Functions // Foundations of Computational Mathematics. 2017. Vol. 17, no. 2. P. 527–566.
21. Shamir O. An Optimal Algorithm for Bandit and Zero-Order Convex Optimization with Two-Point Feedback. // Journal of Machine Learning Research. 2017. Vol. 18, no. 52. P. 1–11.
22. Larson J., Menickelly M., Wild S. M. Derivative-free optimization methods // Acta Numerica. 2019. Vol. 28. P. 287–404.
23. Duchi J. C., Jordan M. I., Wainwright M. J., Wibisono A. Optimal rates for zero-order convex optimization: The power of two function evaluations // IEEE Transactions on Information Theory. 2015. Vol. 61, no. 5. P. 2788–2806.
24. Beznosikov A., Sadiev A., Gasnikov A. Gradient-Free Methods for Saddle-Point Problem. 2020. arXiv:math.OC/2005.05913.
25. Beznosikov A., Gorbunov E., Gasnikov A. Derivative-Free Method For Composite Optimization With Applications To Decentralized Distributed Optimization. 2019.

arXiv:math.OC/1911.10645.

26. Nesterov Y. Introductory Lectures on Convex Optimization: a basic course. Kluwer Academic Publishers, Massachusetts, 2004.
27. Ben-Tal A., Nemirovski A. Lectures on Modern Convex Optimization (Lecture Notes). Personal web-page of A. Nemirovski, 2015.
28. Gorbunov E., Dvurechensky P., Gasnikov A. An accelerated method for derivative-free smooth stochastic convex optimization // arXiv preprint arXiv:1802.09022. 2018.
29. Gasnikov A. Universal gradient descent // MIPT. 2018.
30. Nemirovsky A. S., Yudin D. B. Problem complexity and method efficiency in optimization. 1983.
31. Minsker S. et al. Geometric median and robust estimation in Banach spaces // Bernoulli. 2015. Vol. 21, no. 4. P. 2308–2335.
32. Cohen M. B., Lee Y. T., Miller G. et al. Geometric median in nearly linear time // Proceedings of the forty-eighth annual ACM symposium on Theory of Computing / ACM. 2016. P. 9–21.
33. Chang C.-C., Lin C.-J. LIBSVM: A library for support vector machines // ACM transactions on intelligent systems and technology (TIST). 2011. Vol. 2, no. 3. P. 1–27.