
Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра интеллектуальных систем

Направление подготовки / специальность: 03.03.01 Прикладные математика и физика
Направленность (профиль) подготовки: Математическая физика, компьютерные технологии и
математическое моделирование в экономике

**МЕТОДЫ РЕДУКЦИИ ДИСПЕРСИИ, НЕ ПРЕДПОЛАГАЮЩИЕ
ВЫЧИСЛЕНИЕ ПОЛНОГО ГРАДИЕНТА: ПОВЫШЕНИЕ
ЭФФЕКТИВНОСТИ ЗА СЧЁТ ТЕХНИКИ СЛУЧАЙНОГО
ПЕРЕМЕШИВАНИЯ БАТЧЕЙ**

(бакалаврская работа)

Студент:
Ребриков Алексей Витальевич

(подпись студента)

Научный руководитель:
Безносиков Александр Николаевич,
канд. физ.-мат. наук

(подпись научного руководителя)

Консультант (при наличии):

(подпись консультанта)

Москва 2025

Аннотация

В современном мире машинное обучение невозможно представить без использования больших обучающих выборок и моделей. Это обусловило широкое применение стохастических методов обучения, таких как SGD. Несмотря на простоту, SGD обладает слабыми теоретическими гарантиями сходимости, связанными с неубывающей дисперсией. Данная проблема может быть частично устранена с помощью модификаций, таких как SARAH. Однако эти методы требуют периодического вычисления полного градиента, что может быть затратным по времени. В данной работе были рассмотрены варианты алгоритмов с уменьшением дисперсии, не предполагающие необходимость вычисления полного градиента. Для повышения эффективности по памяти и исключения этих вычислений были использованы два ключевых подхода: эвристика перемешивания и идея, лежащая в основе методов SAG/SAGA. В результате были улучшены существующие оценки для алгоритмов с уменьшением дисперсии без вычисления полного градиента. В случае невыпуклой целевой функции полученная оценка совпадает с классическими методами на основе перемешивания, а для сильно выпуклой задачи достигается улучшение. Проведён всесторонний теоретический анализ, а также представлены масштабные экспериментальные результаты, подтверждающие эффективность и практическую применимость предложенных методов в задачах обучения на больших данных.

Содержание

1 Введение	4
2 Обзор литературы	8
3 Постановка задачи	11
4 Основные результаты	12
4.1 Аппроксимация полного градиента	12
4.2 SARAH без полного градиента	13
4.2.1 Невыпуклая постановка	14
4.2.2 Сильно выпуклая постановка	15
5 Вычислительный эксперимент	16
5.1 Результаты на CIFAR-10 с использованием ResNet-18	17
5.2 Результаты на CIFAR-100 с использованием ResNet-18	18
5.3 Результаты на Tiny ImageNet с использованием Swin Transformer	18
6 Заключение	21
Список литературы	22
Приложение	25

1 Введение

В последние годы в области машинного обучения наблюдается значительный прогресс, обусловленный стремлением к повышению качества решений и возможности решать всё более сложные задачи. Это привело к заметному увеличению как объёма данных, так и масштабов моделей. Данные изменения являются критически важными, поскольку способствуют стабилизации результатов и повышению точности выполнения задач.

Большинство задач машинного обучения сводится к задаче минимизации суммы конечного числа функций:

$$\min_{x \in \mathbb{R}^d} \left[f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x) \right],$$

где $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$, а число функций n велико. Например, в задаче обучения моделей машинного обучения, n соответствует размеру обучающей выборки, а $f_i(x)$ — функции потерь модели на i -м объекте, где x — вектор параметров модели.

Стохастические методы хорошо подходят для данной задачи, так как позволяют избежать вычисления полного градиента на каждой итерации. В условиях реальных задач, где n может быть чрезвычайно велико, такие вычисления становятся крайне ресурсоёмкими. Одним из наиболее известных методов является стохастический градиентный спуск (SGD) и его модификации. На t -й итерации метод выбирает индекс $i_t \in \{1, \dots, n\}$ и выполняет шаг:

$$x^{t+1} = x^t - \gamma \nabla f_{i_t}(x^t),$$

где γ — величина шага метода. При этом $\nabla f_{i_t}(x^t)$ называют *стохастической оценкой градиента* функции f в точке x^t по объекту i_t .

Методы уменьшения дисперсии

Несмотря на простоту SGD, он обладает существенным недостатком: дисперсия стохастических оценок градиента сохраняется большой на протя-

жении всего обучения. В результате, при использовании постоянного шага, метод сходится лишь к окрестности оптимального решения, размер которой зависит от дисперсии. Для решения проблемы высокой дисперсии стохастических градиентных методов были предложены методы уменьшения дисперсии (Variance Reduction, VR), такие как SAG [Schmidt et al., 2017], SAGA [Defazio et al., 2014a], FINITO [Defazio et al., 2014b], SPIDER [Fang et al., 2018] и SARAH [Nguyen et al., 2017]. Эти методы основаны на идее построения более точных оценок градиента, что позволяет существенно улучшить скорость сходимости по сравнению с классическим SGD [Robbins and Monro, 1951]. В частности, метод SARAH использует рекурсивное обновление оценок градиента и обладает как теоретическими, так и практическими преимуществами при оптимизации больших моделей.

Алгоритм 1 SARAH: StochAStic Recursive grAdient algoritHm

Require: начальная точка x^0 , шаг γ , период обновления m

```

1: for  $s = 0, 1, 2, \dots$  do
2:    $v^0 = \nabla f(x^0)$ 
3:    $x^1 = x^0 - \gamma v^0$ 
4:   for  $t = 1$  to  $m$  do
5:     выбрать  $i_t \sim \text{Uniform}(\{1, \dots, n\})$ 
6:      $v^t = \nabla f_{i_t}(x^t) - \nabla f_{i_t}(x^{t-1}) + v^{t-1}$ 
7:      $x^{t+1} = x^t - \gamma v^t$ 
8:   end for
9:    $x^0 \leftarrow x^{m+1}$                                  $\triangleright$  Перезапуск
10: end for

```

В данной работе анализу подвергается метод SARAH (алгоритм 1), в

котором оценка градиента обновляется рекурсивно:

$$\begin{aligned} v^t &= \nabla f_{i_t}(x^t) - \nabla f_{i_t}(x^{t-1}) + v^{t-1}, \\ x^{t+1} &= x^t - \gamma v^t, \end{aligned}$$

где γ — величина шага оптимизации. Основная интуиция данного метода состоит в том, что мы обновляем оценку полного градиента v^t (которая в идеале равна среднему градиенту по всем батчам) разницей градиентов по текущему батчу текущего и предыдущего шага. Конечно, в таком случае это даже не является оценкой полного градиента (так как градиенты по батчам в сумму входят с коэффициентом $\frac{1}{n}$), но тем не менее позволяет уменьшить дисперсию стохастических градиентов, что в свою очередь приводит к более быстрой сходимости метода. Для достижения сходимости к оптимальному решению x^* требуется периодически пересчитывать v^t с использованием полного градиента. Эта процедура выполняется либо через фиксированное число итераций, либо случайным образом. Практическая версия метода, известная как SARAH+ [Nguyen et al., 2017], использует правило автоматического выбора момента обновления, основываясь на величине отношения $\|v^t\|/\|v^0\|$.

Эвристика перемешивания

Важным, но часто недооцениваемым аспектом стохастических методов является способ выбора индексов на каждой итерации. Это напрямую влияет как на стабильность, так и на сходимость. Вместо случайного выбора на каждой итерации, в настоящей работе применяется эвристика перемешивания. Изначально создаётся случайная перестановка индексов $\{1, \dots, n\}$, после чего на каждой итерации используется соответствующий элемент этой перестановки.

Наиболее известные варианты включают:

- **Random Reshuffle (RR)** — данные перемешиваются перед каждой эпохой;
- **Shuffle Once (SO)** — перемешивание только один раз в начале обучения;

- **Cyclic** — данные проходят в фиксированном порядке без перемешивания.

Во всех вариантах перемешивания градиент для каждого объекта вычисляется ровно один раз за эпоху. При этом, выбор по перестановке нарушает свойство несмешённости градиентных оценок:

$$\mathbb{E}_{\pi_s^t} [\nabla f_{\pi_s^t}(x_s^t)] \neq \nabla f(x_s^t),$$

что приводит к более сложному анализу и нестандартным техникам доказательства.

2 Обзор литературы

Методы без вычисления полного градиента

Метод SARAH [Nguyen et al., 2017] на сегодняшний день является одним из стандартных подходов к решению задачи минимизации суммы. Однако, классическая версия требует периодического вычисления полного градиента. В связи с этим был проявлен интерес к вариантам, избегающим таких вычислений.

Методы SAG [Schmidt et al., 2017] и SAGA [Defazio et al., 2014a] решают эту задачу, но требуют хранения дополнительных градиентов, что ведёт к затратам памяти порядка $\mathcal{O}(nd)$, где d — размерность вектора весов, иначе количество параметров, что в современных задачах может достигать десятки миллионов и даже больше. Ряд подходов был предложен для модификации SARAH, исключающей необходимость в полном градиенте.

- В работе [Nguyen et al., 2021] предложен алгоритм inexact-SARAH, в котором полный градиент заменяется на минибатч-оценку (то есть средний градиент по некоторому подмножеству батчей):

$$\frac{1}{|S|} \sum_{i \in S} f_i(x), \quad S \subset \{1, \dots, n\}.$$

- В других работах предлагается гибридная схема без рестартов:

$$v^t = \beta_t \nabla f_{i_t}(x^t) + (1 - \beta_t)(\nabla f_{i_t}(x^t) - \nabla f_{i_t}(x^{t-1}) + v^{t-1}),$$

где параметр β_t либо постоянен [Liu et al., 2020], либо стремится к нулю, как в методе STORM [Cutkosky and Orabona, 2019]. Метод ZEROSARAH [Li et al., 2021] сочетает такую схему с SAG/SAGA.

Методы с перемешиванием

Так как рассматриваемый метод использует эвристику перемешивания, необходимо проанализировать существующие подходы. При выборке без воз-

вращения к уже учтенным батчам на протяжении эпохи каждый объект используется ровно один раз. Было показано, что Random Reshuffle может сходиться быстрее, чем SGD, на практике.

Однако теоретические оценки долгое время отставали от классических методов с независимым выбором индексов. Прорыв был достигнут в работе [Mishchenko et al., 2020], где представлены новые методы анализа. Для сильно выпуклых задач получены такие же оценки, как у SGD с независимым выбором. Однако в невыпуклом случае результаты остались слабее, а также требовали большого числа эпох, что не характерно для современных нейронных сетей. Альтернатива была предложена в [Koloskova et al., 2024], где анализ строится на так называемом «периоде корреляции» вместо полной эпохи.

Позднее эвристика перемешивания была применена к более общим задачам вариационных неравенств и, в частности, к методу EXTRAGRADIENT. Это позволило получить аналогичные линейные оценки. В дальнейшем внимание было сосредоточено на сочетании перемешивания с методами уменьшения дисперсии.

Таблица 1: Сравнение оценок сходимости различных алгоритмов. Красным цветом выделены оценки, у которых выигрывает алгоритм SARAH из данной работы.

Алгоритм	Без полного градиента?	Память	Невыпуклая задача	Сильно выпуклая задача
SAGA [Park and Ryu, 2020]	✓	$\mathcal{O}(nd)$	\	$\mathcal{O}\left(n \frac{L^2}{\mu^2} \log\left(\frac{1}{\varepsilon}\right)\right)$
IAG [Gurbuzbalaban et al., 2017]	✓	$\mathcal{O}(nd)$	\	$\mathcal{O}\left(n^{\frac{2}{3}} \frac{L^2}{\mu^2} \log\left(\frac{1}{\varepsilon}\right)\right)$
PIAG [Vanli et al., 2016]	✓	$\mathcal{O}(nd)$	\	$\mathcal{O}\left(n \frac{L}{\mu} \log\left(\frac{1}{\varepsilon}\right)\right)$
DIAG [Mokhtari et al., 2018]	✓	$\mathcal{O}(nd)$	\	$\mathcal{O}\left(n \frac{L}{\mu} \log\left(\frac{1}{\varepsilon}\right)\right)$
Prox-DFinito [Huang et al., 2021]	✓	$\mathcal{O}(nd)$	\	$\mathcal{O}\left(n \frac{L}{\mu} \log\left(\frac{1}{\varepsilon}\right)\right)$
AVRG [Ying et al., 2020]	✓	$\mathcal{O}(d)$	\	$\mathcal{O}\left(n \frac{L^2}{\mu^2} \log\left(\frac{1}{\varepsilon}\right)\right)$
SVRG [Sun et al., 2019]	✗	$\mathcal{O}(d)$	\	$\mathcal{O}\left(n^{\frac{3}{2}} \frac{L^2}{\mu^2} \log\left(\frac{1}{\varepsilon}\right)\right)$
SVRG [Malinovsky et al., 2023]	✗	$\mathcal{O}(d)$	$\mathcal{O}\left(\frac{nL}{\varepsilon^2}\right)$	$\mathcal{O}\left(n^{\frac{3}{2}} \frac{L^{3/2}}{\mu^{3/2}} \log\left(\frac{1}{\varepsilon}\right)\right)^{(1)}$
SARAH [Beznosikov and Takáč, 2023]	✓	$\mathcal{O}(d)$	\	$\mathcal{O}\left(n^{\frac{2}{3}} \frac{L}{\mu} \log\left(\frac{1}{\varepsilon}\right)\right)$
SARAH (данная работа)	✓	$\mathcal{O}(d)$	$\mathcal{O}\left(\frac{nL}{\varepsilon^2}\right)$	$\mathcal{O}\left(n \frac{L}{\mu} \log\left(\frac{1}{\varepsilon}\right)\right)$

Столбцы: «Без полного градиента?» — указывает, требует ли метод вычисления полного градиента. «Память» — дополнительные затраты по памяти.

Обозначения: μ — коэффициент сильной выпуклости, L — константа гладкости, n — размер выборки, d — размерность задачи, ε — требуемая точность.

(1) В данной работе также получены улучшенные оценки в случае больших данных: $n \gg \mathcal{O}\left(\frac{L}{\mu}\right)$, однако они выходят за рамки настоящего анализа.

3 Постановка задачи

В данной работе рассматривается задача оптимизации конечной суммы функций следующего вида:

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) \right], \quad (1)$$

где каждая функция $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ соответствует функции потерь на одном элементе выборки. Предполагается, что f дифференцируема, однако может быть как выпуклой, так и невыпуклой. Основная цель — нахождение точки, минимизирующей функцию f с использованием стохастических градиентных методов, не требующих вычисления полного градиента на каждой итерации.

Для теоретического анализа вводится ряд стандартных предположений.

Предположение 1 (Гладкость функций). *Каждая функция f_i обладает L -гладкостью, то есть для любых $x, y \in \mathbb{R}^d$ выполняется неравенство:*

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

Предположение 2 (Сильная выпуклость). *Каждая функция f_i является μ -сильно выпуклой, то есть для любых $x, y \in \mathbb{R}^d$ выполняется:*

$$f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2.$$

Предположение 3 (Невыпуклость). *Функция f может быть невыпуклой, но при этом обладает конечным инфимумом:*

$$f^* := \inf_{x \in \mathbb{R}^d} f(x) > -\infty.$$

4 Основные результаты

4.1 Аппроксимация полного градиента

В данном разделе представлена эвристика, основанная на перемешивании выборки и идеях SAG [Schmidt et al., 2017] и SAGA [Defazio et al., 2014a], позволяющая аппроксимировать полный градиент без необходимости хранения всех предыдущих значений градиентов. Алгоритм SAG был одним из первых методов, направленных на ускорение стохастического градиентного спуска за счёт уменьшения дисперсии обновлений. В SAG обновление имеет вид:

$$x^{t+1} = x^t - \frac{\gamma}{n} \left(\nabla f_{i_t}(x^t) - \nabla f_{i_t}(\phi_{i_t}^t) + \sum_{j=1}^n \nabla f_j(\phi_j^t) \right), \quad (2)$$

где ϕ_j^t — точка, в которой ранее был вычислен градиент функции f_j . Например, после такого шага, $\phi_{i_t}^t = x^t$ и $\phi_j^t = \phi_j^{t-1}$ для всех $j \neq i_t$. В данном подходе на каждом шаге обновляется один из элементов суммы, что снижает дисперсию оценки градиента.

При случайной выборке i_t в SAG сложно отследить, когда последний раз обновлялся градиент для конкретного индекса. Однако при использовании перемешивания (shuffling) известно, что в течение эпохи все градиенты ∇f_j будут вычислены. Таким образом, в начале каждой эпохи возможна аппроксимация полного градиента:

$$v_{s+1} = \frac{1}{n} \sum_{t=1}^n \nabla f_{\pi_s^t}(x_s^t), \quad (3)$$

где π_s^t — перестановка индексов после перемешивания в начале эпохи s . Стоит отметить, что это будет не точная оценка, так как градиенты по батчам считаются в разных точках. При этом расчёт может быть реализован через скользящее среднее без дополнительных затрат памяти:

$$\tilde{v}_s^0 = 0, \quad \tilde{v}_s^{t+1} = \frac{t}{t+1} \tilde{v}_s^t + \frac{1}{t+1} \nabla f_{\pi_s^t}(x_s^t), \quad v_{s+1} = \tilde{v}_s^n. \quad (4)$$

Лемма 1. Формулы (3) и (4) эквивалентны.

4.2 SARAH без полного градиента

Алгоритм SARAH зарекомендовал себя как эффективный метод снижения дисперсии градиентных оценок, обладающий практическими преимуществами по сравнению с альтернативами. В данном разделе рассматривается модификация алгоритма, исключающая необходимость пересчёта полного градиента. Ниже представлена формальная постановка алгоритма.

Алгоритм 2 No FULL GRAD SARAH

- 1: **Вход:** Начальное приближение $x_0^0 \in \mathbb{R}^d$; Начальные градиенты $\tilde{v}_0^0 = 0^d, v_0 = 0^d$
 - 2: **Параметр:** Шаг градиентного спуска $\gamma > 0$
 - 3: **for** эпохи $s = 0, 1, 2, \dots, S$ **do**
 - 4: Сэмплируется перестановка π_s^1, \dots, π_s^n из $\overline{1, n}$ ▷ по эвристике
 - 5: $v_s^0 = v_s$
 - 6: $x_s^1 = x_s^0 - \gamma v_s^0$
 - 7: **for** $t = 1, 2, \dots, n$ **do**
 - 8: $\tilde{v}_s^{t+1} = \frac{t-1}{t}\tilde{v}_s^t + \frac{1}{t}\nabla f_{\pi_s^t}(x_s^t)$
 - 9: $v_s^t = \frac{1}{n}(\nabla f_{\pi_s^t}(x_s^t) - \nabla f_{\pi_s^t}(x_s^{t-1})) + v_s^{t-1}$
 - 10: $x_s^{t+1} = x_s^t - \gamma v_s^t$
 - 11: **end for**
 - 12: $x_{s+1}^0 = x_s^{n+1}$
 - 13: $\tilde{v}_{s+1}^1 = 0$
 - 14: $v_{s+1} = \tilde{v}_s^{n+1}$
 - 15: **end for**
-

Модификация алгоритма позволяет отказаться от пересчёта полного градиента, используя идею скользящего усреднения стохастических градиентов. Обновление в строке 8 учитывает изменение индексации: усреднение начинает-

ся с $t = 1$, а не с $t = 0$, что позволяет избежать появления лишнего множителя $\frac{1}{n+1}$ в оценках.

4.2.1 Невыпуклая постановка

Для более детального анализа метода следует рассмотреть промежуточные результаты. Анализ структурируется следующим образом: сначала изучается сходимость в пределах одной эпохи, после чего полученные оценки распространяются рекурсивно на все эпохи. Ключевым моментом является выявление характера изменений градиентов от начала эпохи до различных её точек. Необходимо установить, что данные изменения обусловлены двумя основными факторами: точностью приближения полного градиента в начале эпохи и степенью отклонения обновлений от исходной точки в процессе прохождения эпохи. Для обоснования данного утверждения формулируется лемма.

Лемма 2. *Пусть выполнены предположения 1, 2, 3. Тогда для алгоритма 2 справедлива оценка:*

$$\left\| \nabla f(x_s^0) - \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \leq 2\|\nabla f(x_s^0) - v_s\|^2 + \frac{2L^2}{n+1} \sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2.$$

Полученное выражение содержит два слагаемых: первое отвечает за точность аппроксимации полного градиента, второе — за накопленную ошибку, возникающую при отклонении траектории оптимизации от начальной точки. Далее доказывается следующая лемма:

Лемма 3. *Пусть выполнены предположения 1, 2, 3 и шаг $\gamma \leq \frac{1}{3L}$. Тогда для алгоритма 2 справедливо:*

$$\left\| \nabla f(x_s^0) - \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \leq 9\gamma^2 L^2 \|v_s\|^2 + 36\gamma^2 L^2 n^2 \|v_{s-1}\|^2.$$

На основании этих лемм формулируется основная теорема:

Теорема 1. Пусть выполнены предположения 1, 3. Тогда алгоритму 2 при шаге $\gamma \leq \frac{1}{20L(n+1)}$ для достижения ε -точности, определяемой как $\varepsilon^2 = \frac{1}{S} \sum_{s=1}^S \|\nabla f(x_s^0)\|^2$, требуется

$$\mathcal{O}\left(\frac{nL}{\varepsilon^2}\right) \text{ итераций и вызовов оракула.}$$

4.2.2 Сильно выпуклая постановка

Переход к сильно выпуклой постановке осуществляется с использованием условия Поляка-Лоясевича (см. Приложение А).

Теорема 2. Пусть выполнены предположения 1, 2. Тогда алгоритму 2 при шаге $\gamma \leq \frac{1}{20L(n+1)}$ для достижения ε -точности, определяемой как $\varepsilon = f(x_{S+1}^0) - f(x^*)$, требуется

$$\mathcal{O}\left(\frac{nL}{\mu} \log \frac{1}{\varepsilon}\right) \text{ итераций и вызовов оракула.}$$

Полученные результаты для алгоритма NO FULL GRAD SARAH в условиях сильной выпуклости аналогичны тем, что наблюдаются в невыпуклом случае. Более того, полученные оценки значительно превосходят существующие на данный момент результаты для методов без вычисления полного градиента. Сравнение с другими методами на основе перемешивания (см. Таблицу 1) показывает, что предложенный алгоритм улучшает гарантии сходимости при сохранении оптимального объёма дополнительной памяти. Таким образом, он вносит вклад в развитие всего класса алгоритмов на основе перемешивания.

5 Вычислительный эксперимент

Целью данного раздела является эмпирическая проверка эффективности предложенного алгоритма NO FULL GRAD SARAH на задачах классификации изображений. Были проведены вычислительные эксперименты на следующих датасетах:

- CIFAR-10 и CIFAR-100 с использованием архитектуры ResNet-18;
- TINY IMAGENET с использованием модели Swin Transformer.

Во всех экспериментах обучение осуществлялось с размером батча 128. Параметр регуляризации веса был установлен как $\lambda_1 = 5 \times 10^{-4}$. Для каждой модели фиксировались метрики качества на обучающей и тестовой выборках: кросс-энтропийная функция потерь и точность. Метрики визуализировались в зависимости от числа эквивалентных вызовов полного градиента.

5.1 Результаты на CIFAR-10 с использованием ResNet-18

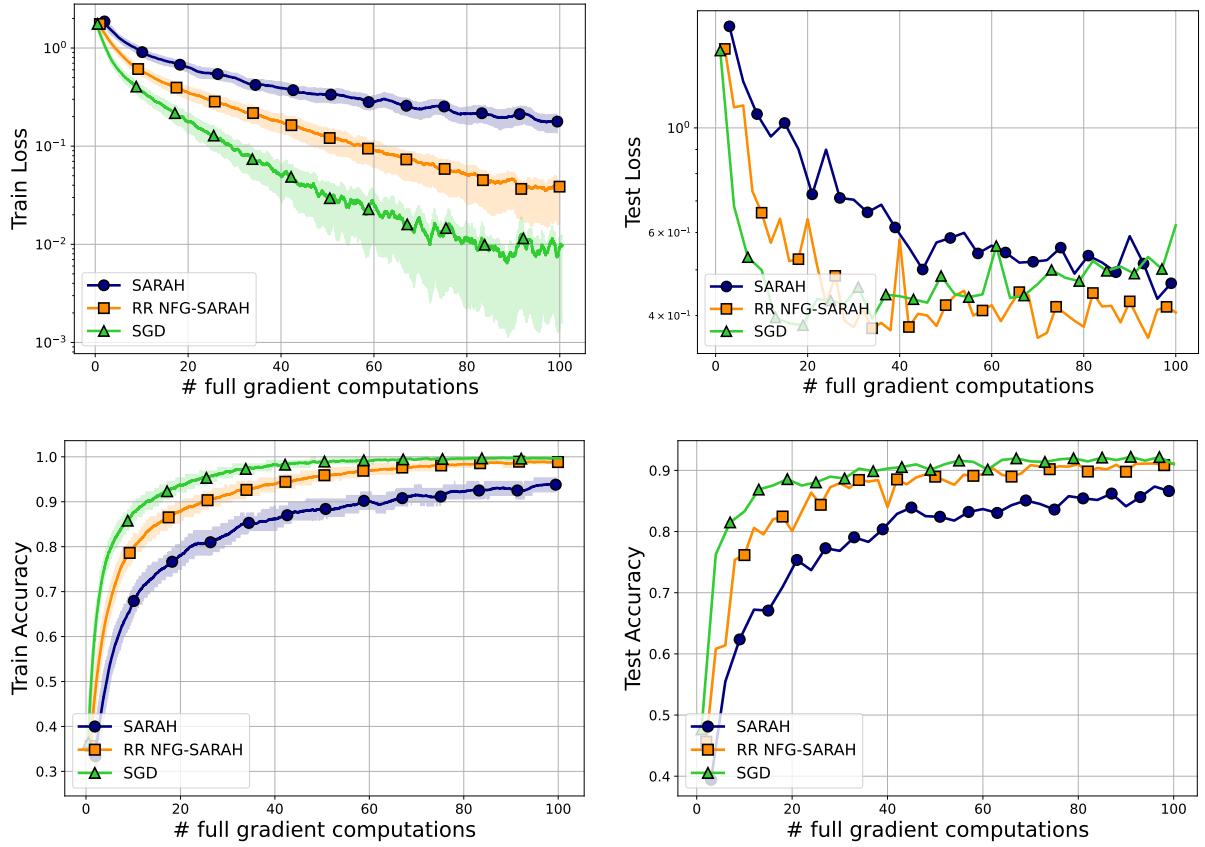


Рис. 1: Сходимость NO FULL GRAD SARAH и классического SARAH на датасете CIFAR-10.

Полученные результаты демонстрируют стабильное убывание функции потерь при обучении методом NO FULL GRAD SARAH, опережая оригинальный алгоритм по скорости сходимости при одинаковом количестве эквивалентных вызовов полного градиента. На тестовой выборке также наблюдается уменьшение функции потерь до более низкого значения, при этом финальная точность постепенно улучшается и достигает более высоких значений на поздних стадиях обучения.

5.2 Результаты на CIFAR-100 с использованием ResNet-18

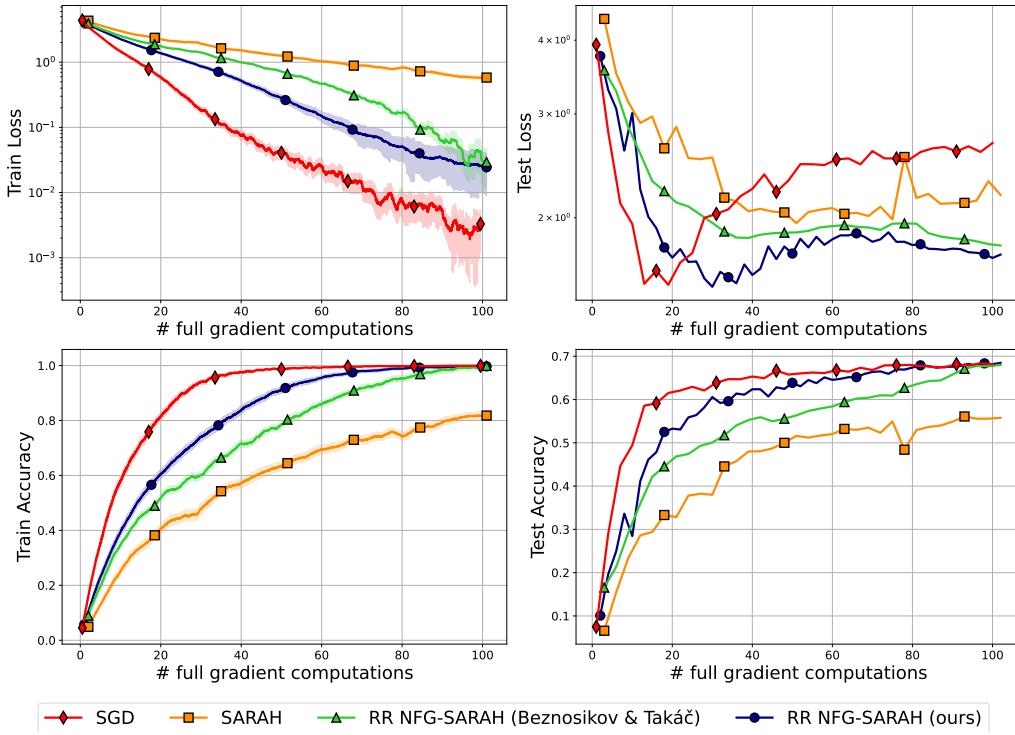


Рис. 2: Сходимость NO FULL GRAD SARAH и SARAH на датасете CIFAR-100.

На CIFAR-100 также была получена устойчивая сходимость предложенного метода. Тестовая ошибка продолжает уменьшаться, даже после того как классический метод достигает минимума. Отметим также, что точность возрастает значительно быстрее на поздних стадиях обучения, что указывает на улучшенное обобщающее поведение.

5.3 Результаты на Tiny ImageNet с использованием Swin Transformer

Был использован датасет Tiny ImageNet, включающий 200 классов изображений размером 64×64 , увеличенных до 224×224 с целью соответствия входу модели. Архитектура — Tiny Swin Transformer (`swin_T_patch4_window7_224`),

инициализированная предобученными весами с ImageNet-1K. Обучение велось с градиентным клиппингом на уровне 1.0. Размер батча — 256.

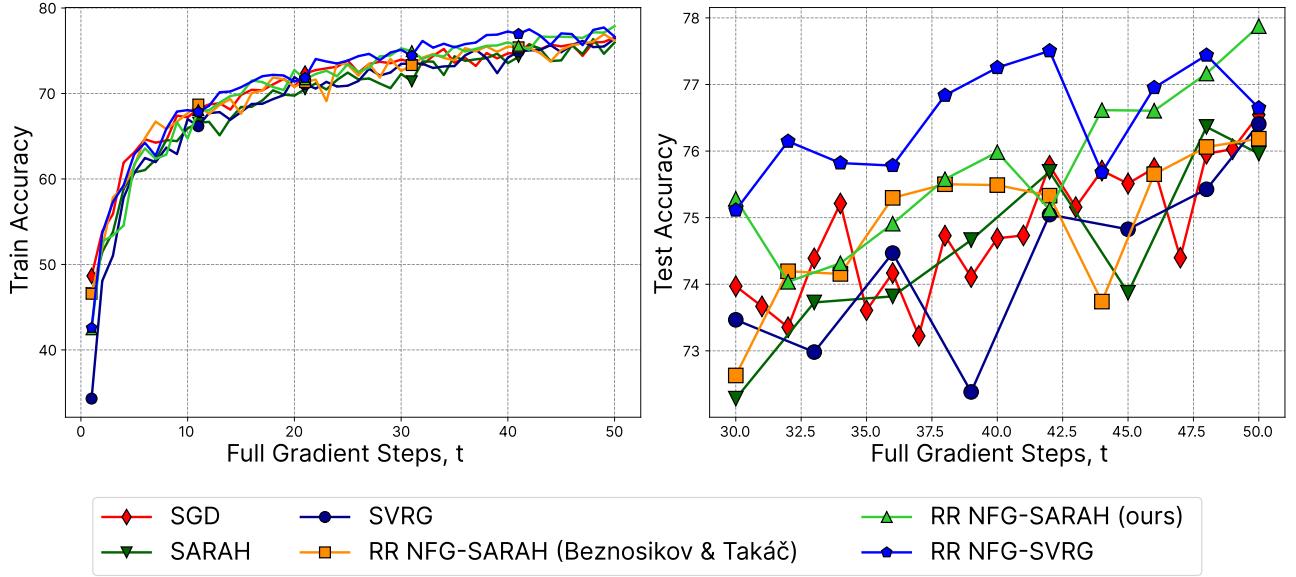


Рис. 3: Сходимость NO FULL GRAD SARAH и других методов на Tiny ImageNet.

Таблица 2: Финальная точность различных методов на Tiny ImageNet.

Метод	Точность (\uparrow)
SGD	76.545%
SARAH	75.961%
RR NFG-SARAH (по [Beznosikov and Takáč, 2023])	76.186%
RR NFG-SARAH (предложенный)	77.875%

Модифицированный алгоритм NO FULL GRAD SARAH, реализованный в рамках данной работы, демонстрирует превосходство над как классическими,

так и ранее предложенными модифицированными алгоритмами. Особенно заметно улучшение в задачах с большим числом параметров и высокой сложностью модели, таких как Swin Transformer.

6 Заключение

В данной работе был проведён анализ метода стохастического градиентного типа без полного градиента, основанного на итеративном уточнении оценок градиента. Рассмотренный алгоритм обладает простой структурой обновления.

Были установлены оценки скорости сходимости в двух постановках: невыпуклой и сильно выпуклой. В первом случае подтверждающая эффективность метода при минимальных предположениях о функции. Во втором случае соответствующая классическим результатам для методов первого порядка в сильно выпуклых задачах.

Кроме того, для теоретического обоснования приведены ключевые леммы, использованные при получении оценок. Все выводы сопровождаются строгими доказательствами, а ограничения на шаг метода подобраны таким образом, чтобы обеспечить выполнение условий сходимости при минимальных требованиях к гиперпараметрам.

Полученные результаты подтверждают практическую применимость алгоритма и могут быть использованы в дальнейшем при разработке более сложных схем стохастической оптимизации с пониженной вычислительной сложностью.

Список литературы

- Aleksandr Beznosikov and Martin Takáč. Random-reshuffled sarah does not need full gradient computations. *Optimization Letters*, pages 1–23, 2023.
- Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.
- Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014a.
- Aaron Defazio, Justin Domke, et al. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning*, pages 1125–1133. PMLR, 2014b.
- Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.
- Mert Gurbuzbalaban, Asuman Ozdaglar, and Pablo A Parrilo. On the convergence rate of incremental aggregated gradient algorithms. *SIAM Journal on Optimization*, 27(2):1035–1048, 2017.
- Xinmeng Huang, Kun Yuan, Xianghui Mao, and Wotao Yin. An improved analysis and rates for variance reduction under without-replacement sampling orders. *Advances in Neural Information Processing Systems*, 34:3232–3243, 2021.
- Anastasia Koloskova, Nikita Doikov, Sebastian U. Stich, and Martin Jaggi. On convergence of incremental gradient for non-convex smooth functions, 2024.
- Zhize Li, Slavomír Hanzely, and Peter Richtárik. Zerosarah: Efficient nonconvex finite-sum optimization with zero full gradient computation. *arXiv preprint arXiv:2103.01447*, 2021.

Deyi Liu, Lam M Nguyen, and Quoc Tran-Dinh. An optimal hybrid variance-reduced algorithm for stochastic composite nonconvex optimization. *arXiv preprint arXiv:2008.09055*, 2020.

Grigory Malinovsky, Alibek Sailanbayev, and Peter Richtárik. Random reshuffling with variance reduction: New analysis and better rates. In *Uncertainty in Artificial Intelligence*, pages 1347–1357. PMLR, 2023.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020.

Aryan Mokhtari, Mert Gurbuzbalaban, and Alejandro Ribeiro. Surpassing gradient descent provably: A cyclic incremental method with linear convergence rate. *SIAM Journal on Optimization*, 28(2):1420–1447, 2018.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pages 2613–2621. PMLR, 2017.

Lam M Nguyen, Katya Scheinberg, and Martin Takáč. Inexact sarah algorithm for stochastic optimization. *Optimization Methods and Software*, 36(1):237–258, 2021.

Youngsuk Park and Ernest K Ryu. Linear convergence of cyclic saga. *Optimization Letters*, 14(6):1583–1598, 2020.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

Tao Sun, Yuejiao Sun, Dongsheng Li, and Qing Liao. General proximal incremental aggregated gradient algorithms: Better and novel results under general scheme. *Advances in Neural Information Processing Systems*, 32, 2019.

Nuri Denizcan Vanli, Mert Gurbuzbalaban, and Asu Ozdaglar. A stronger convergence result on the proximal incremental aggregated gradient method. *arXiv preprint arXiv:1611.08022*, 2016.

Bicheng Ying, Kun Yuan, and Ali H Sayed. Variance-reduced stochastic learning under random reshuffling. *IEEE Transactions on Signal Processing*, 68:1390–1408, 2020.

Приложение

A Основные неравенства

В данном разделе формулируются неравенства, использующиеся в дальнейших оценках. Пусть функция f удовлетворяет условию L -гладкости (см. Предположение 1), а функция g — условию μ -сильной выпуклости (см. Предположение 2). Тогда для любых векторов $x, y, \{x_i\} \subset \mathbb{R}^d$ и положительных скаляров $\alpha, \beta > 0$ справедливы следующие утверждения.

$$2\langle x, y \rangle \leq \frac{\|x\|^2}{\alpha} + \alpha\|y\|^2, \quad (\text{Scalar})$$

$$2\langle x, y \rangle = \|x + y\|^2 - \|x\|^2 - \|y\|^2, \quad (\text{Norm})$$

$$\|x + y\|^2 \leq (1 + \beta)\|x\|^2 + \left(1 + \frac{1}{\beta}\right)\|y\|^2, \quad (\text{Quad})$$

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2, \quad (\text{Lip})$$

$$\left\| \sum_{i=1}^n x_i \right\|^2 \leq n \sum_{i=1}^n \|x_i\|^2, \quad (\text{CS})$$

$$g(x) - \inf g \leq \frac{1}{2\mu} \|\nabla g(x)\|^2. \quad (\text{PL})$$

Лемма 4 (Неравенство для скалярного произведения). Для любых $x, y \in \mathbb{R}^d$ и $\alpha > 0$ выполнено:

$$2\langle x, y \rangle \leq \frac{\|x\|^2}{\alpha} + \alpha\|y\|^2. \quad (\text{Scalar})$$

Доказательство. Рассмотрим скалярное произведение $\langle x, y \rangle$. В силу неравенства Коши–Буняковского имеем:

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|.$$

При этом для любых $a, b \in \mathbb{R}_{\geq 0}$ и $\alpha > 0$ справедливо неравенство между средним арифметическим и средним геометрическим:

$$2ab = 2\sqrt{\frac{a^2}{\alpha} \cdot \alpha b^2} \leq \frac{a^2}{\alpha} + \alpha b^2.$$

Применяя это неравенство к $a = \|x\|$, $b = \|y\|$, получаем:

$$2\|x\| \cdot \|y\| \leq \frac{\|x\|^2}{\alpha} + \alpha\|y\|^2.$$

Поскольку $\langle x, y \rangle \leq \|x\| \cdot \|y\|$, заключаем:

$$2\langle x, y \rangle \leq 2\|x\| \cdot \|y\| \leq \frac{\|x\|^2}{\alpha} + \alpha\|y\|^2.$$

Таким образом, утверждение доказано. \square

Лемма 5 (Тождество параллелограмма). Для любых $x, y \in \mathbb{R}^d$ выполнено:

$$2\langle x, y \rangle = \|x + y\|^2 - \|x\|^2 - \|y\|^2. \quad (\text{Norm})$$

Доказательство. Рассмотрим квадрат нормы суммы двух векторов:

$$\|x + y\|^2 = \langle x + y, x + y \rangle.$$

Путём раскрытия скобок получено:

$$\langle x + y, x + y \rangle = \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle.$$

Поскольку $\langle x, y \rangle = \langle y, x \rangle$, это выражение упрощается до:

$$\|x + y\|^2 = \|x\|^2 + 2\langle x, y \rangle + \|y\|^2.$$

Перенося $\|x\|^2 + \|y\|^2$ в левую часть, получаем утверждение леммы:

$$2\langle x, y \rangle = \|x + y\|^2 - \|x\|^2 - \|y\|^2.$$

\square

Лемма 6 (Квадратичное неравенство). Для любых $x, y \in \mathbb{R}^d$ и $\beta > 0$ выполнено:

$$\|x + y\|^2 \leq (1 + \beta)\|x\|^2 + \left(1 + \frac{1}{\beta}\right)\|y\|^2. \quad (\text{Quad})$$

Доказательство. Рассмотрим выражение $\|x + y\|^2$. Согласно тождеству леммы 5, имеем:

$$\|x + y\|^2 = \|x\|^2 + 2\langle x, y \rangle + \|y\|^2.$$

Применим неравенство из леммы 4 (при $\alpha = \beta$) к скалярному произведению:

$$2\langle x, y \rangle \leq \frac{\|x\|^2}{\beta} + \beta\|y\|^2.$$

Подставляя это в выражение для $\|x + y\|^2$, получаем:

$$\begin{aligned}\|x + y\|^2 &\leq \|x\|^2 + \frac{\|x\|^2}{\beta} + \beta\|y\|^2 + \|y\|^2 \\ &= \left(1 + \frac{1}{\beta}\right)\|x\|^2 + (1 + \beta)\|y\|^2.\end{aligned}$$

Переобозначив $\beta \mapsto \frac{1}{\beta}$, получаем утверждение леммы:

$$\|x + y\|^2 \leq (1 + \beta)\|x\|^2 + \left(1 + \frac{1}{\beta}\right)\|y\|^2.$$

□

Лемма 7 (Гладкость функции f). *Если функция f удовлетворяет условию L -гладкости, то для любых $x, y \in \mathbb{R}^d$ выполнено:*

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2. \quad (\text{Lip})$$

Доказательство. Пусть $f : \mathbb{R}^d \rightarrow \mathbb{R}$ — дифференцируемая функция с L -липшицевым градиентом, то есть для любых $u, v \in \mathbb{R}^d$ выполнено:

$$\|\nabla f(u) - \nabla f(v)\| \leq L\|u - v\|.$$

Рассмотрим вспомогательную функцию $\varphi : [0, 1] \rightarrow \mathbb{R}$, заданную как:

$$\varphi(t) = f(y + t(x - y)).$$

Тогда $\varphi(0) = f(y)$, $\varphi(1) = f(x)$, и по формуле производной по цепному правилу:

$$\varphi'(t) = \langle \nabla f(y + t(x - y)), x - y \rangle.$$

Воспользуемся формулой Ньютона–Лейбница:

$$f(x) - f(y) = \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t) dt = \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt.$$

Поскольку $\nabla f(y + t(x - y)) = \nabla f(y) + [\nabla f(y + t(x - y)) - \nabla f(y)]$, выражение под интегралом можно представить как:

$$\langle \nabla f(y), x - y \rangle + \langle \nabla f(y + t(x - y)) - \nabla f(y), x - y \rangle.$$

Следовательно:

$$\begin{aligned} f(x) - f(y) &= \langle \nabla f(y), x - y \rangle + \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), x - y \rangle dt \\ &\leq \langle \nabla f(y), x - y \rangle + \int_0^1 \|\nabla f(y + t(x - y)) - \nabla f(y)\| \cdot \|x - y\| dt \\ &\leq \langle \nabla f(y), x - y \rangle + \int_0^1 Lt \|x - y\|^2 dt \\ &= \langle \nabla f(y), x - y \rangle + L \|x - y\|^2 \int_0^1 t dt \\ &= \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \end{aligned}$$

что и доказывает утверждение. \square

Лемма 8 (Обобщённое неравенство Коши–Буняковского–Шварца). Для любых $x_1, \dots, x_n \in \mathbb{R}^d$ выполнено:

$$\left\| \sum_{i=1}^n x_i \right\|^2 \leq n \sum_{i=1}^n \|x_i\|^2. \quad (\text{CS})$$

Доказательство. Обозначим $S = \sum_{i=1}^n x_i$. Тогда

$$\|S\|^2 = \left\langle \sum_{i=1}^n x_i, \sum_{j=1}^n x_j \right\rangle = \sum_{i=1}^n \sum_{j=1}^n \langle x_i, x_j \rangle.$$

Применим неравенство Коши–Шварца для скалярного произведения:

$$\langle x_i, x_j \rangle \leq \|x_i\| \cdot \|x_j\|.$$

Следовательно:

$$\|S\|^2 \leq \sum_{i=1}^n \sum_{j=1}^n \|x_i\| \cdot \|x_j\|.$$

Рассмотрим полную сумму:

$$\sum_{i=1}^n \sum_{j=1}^n \|x_i\| \cdot \|x_j\| = \left(\sum_{i=1}^n \|x_i\| \right)^2 \leq n \sum_{i=1}^n \|x_i\|^2,$$

где использовано неравенство между средним арифметическим и средним квадратичным применительно к вектору из норм $\|x_1\|, \dots, \|x_n\|$.

Так как

$$\|S\|^2 \leq \sum_{i=1}^n \sum_{j=1}^n \|x_i\| \cdot \|x_j\| \leq n \sum_{i=1}^n \|x_i\|^2,$$

утверждение доказано. \square

Лемма 9 (PL-условие для g). *Пусть функция $g : \mathbb{R}^d \rightarrow \mathbb{R}$ μ -сильно выпукла с константой $\mu > 0$ и дифференцируема. Тогда для любых $x \in \mathbb{R}^d$ выполнено:*

$$g(x) - \inf g \leq \frac{1}{2\mu} \|\nabla g(x)\|^2. \quad (\text{PL})$$

Доказательство. Обозначим через $x^* \in \arg \min g$ точку глобального минимума функции g , тогда $\nabla g(x^*) = 0$ и $g(x^*) = \inf g$. Из определения μ -сильной выпуклости следует, что для любых $x \in \mathbb{R}^d$ выполнено:

$$g(x^*) \geq g(x) + \langle \nabla g(x), x^* - x \rangle + \frac{\mu}{2} \|x^* - x\|^2.$$

Преобразуем неравенство:

$$g(x) - g(x^*) \leq \langle \nabla g(x), x - x^* \rangle - \frac{\mu}{2} \|x - x^*\|^2.$$

Для оценки скалярного произведения применим неравенство (Scalar):

$$\langle \nabla g(x), x - x^* \rangle \leq \frac{1}{2\mu} \|\nabla g(x)\|^2 + \frac{\mu}{2} \|x - x^*\|^2.$$

Тогда:

$$g(x) - g(x^*) \leq \left(\frac{1}{2\mu} \|\nabla g(x)\|^2 + \frac{\mu}{2} \|x - x^*\|^2 \right) - \frac{\mu}{2} \|x - x^*\|^2 = \frac{1}{2\mu} \|\nabla g(x)\|^2,$$

что и доказывает утверждение. \square

В Доказательства утверждений

Для удобства изложения кратко описывается алгоритм 2. В случае, когда рассматривается эпоха $s \neq 0$, правило обновления принимает следующий вид:

$$\left\{ \begin{array}{l} \text{при } t = 0: \\ x_s^0 = x_{s-1}^n, \\ v_s^0 = v_s = \frac{1}{n} \sum_{t=1}^n \nabla f_{\pi_{s-1}^t}(x_{s-1}^t), \\ x_s^1 = x_s^0 - \gamma v_s^0; \\ \text{на последующих итерациях в пределах эпохи:} \\ v_s^t = v_s^{t-1} + \frac{1}{n} (\nabla f_{\pi_s^t}(x_s^t) - \nabla f_{\pi_s^t}(\omega_s)), \\ x_s^{t+1} = x_s^t - \gamma v_s^t. \end{array} \right. \quad (5)$$

B.1 Вспомогательные леммы

Сперва будет доказана лемма про подсчет оценки полного градиента через скользящее среднее.

Лемма 10 (Лемма 1). *Формулы*

$$v_{s+1} = \frac{1}{n} \sum_{t=1}^n \nabla f_{\pi_s^t}(x_s^t), \quad (6)$$

где π_s^t — перестановка индексов после перемешивания в начале эпохи s , и

$$\tilde{v}_s^0 = 0, \quad \tilde{v}_s^{t+1} = \frac{t}{t+1} \tilde{v}_s^t + \frac{1}{t+1} \nabla f_{\pi_s^t}(x_s^t), \quad v_{s+1} = \tilde{v}_s^n. \quad (7)$$

эквивалентны.

Доказательство. Будет показано, что для любого $t \in \{1, \dots, n\}$ выполняется представление:

$$\tilde{v}_s^t = \frac{1}{t} \sum_{i=1}^t \nabla f_{\pi_s^i}(x_s^i).$$

База индукции. При $t = 1$ с использованием инициализации $\tilde{v}_s^0 = 0$ из (7) получено:

$$\tilde{v}_s^1 = \frac{0}{1} \cdot 0 + \frac{1}{1} \nabla f_{\pi_s^1}(x_s^1) = \nabla f_{\pi_s^1}(x_s^1),$$

что совпадает с $\frac{1}{1} \sum_{i=1}^1 \nabla f_{\pi_s^i}(x_s^i)$.

Переход. Пусть для некоторого $t \geq 1$ выполнено:

$$\tilde{v}_s^t = \frac{1}{t} \sum_{i=1}^t \nabla f_{\pi_s^i}(x_s^i).$$

Тогда, подставляя в формулу (7), получено:

$$\begin{aligned} \tilde{v}_s^{t+1} &= \frac{t}{t+1} \tilde{v}_s^t + \frac{1}{t+1} \nabla f_{\pi_s^{t+1}}(x_s^{t+1}) \\ &= \frac{t}{t+1} \cdot \left(\frac{1}{t} \sum_{i=1}^t \nabla f_{\pi_s^i}(x_s^i) \right) + \frac{1}{t+1} \nabla f_{\pi_s^{t+1}}(x_s^{t+1}) \\ &= \frac{1}{t+1} \sum_{i=1}^t \nabla f_{\pi_s^i}(x_s^i) + \frac{1}{t+1} \nabla f_{\pi_s^{t+1}}(x_s^{t+1}) \\ &= \frac{1}{t+1} \sum_{i=1}^{t+1} \nabla f_{\pi_s^i}(x_s^i), \end{aligned}$$

что завершает индукционный переход.

Следовательно, после n шагов рекурсии выполняется равенство:

$$\tilde{v}_s^n = \frac{1}{n} \sum_{i=1}^n \nabla f_{\pi_s^i}(x_s^i) = v_{s+1},$$

что и доказывает эквивалентность выражений (6) и (7). \square

B.2 Невыпуклый случай

Лемма 11. Пусть выполняются Предположения 1 и 3. Если шаг градиентного спуска удовлетворяет условию $\gamma \leq \frac{1}{L(n+1)}$, то для алгоритма 2 выполняется неравенство

$$f(x_{s+1}^0) \leq f(x_s^0) - \frac{\gamma(n+1)}{2} \|\nabla f(x_s^0)\|^2 + \frac{\gamma(n+1)}{2} \left\| \nabla f(x_s^0) - \frac{1}{n+1} \sum_{i=0}^n v_s^i \right\|^2.$$

Доказательство. На основе итерационной схемы алгоритма 2, заданной в (5), получаем:

$$\begin{aligned}
f(x_{s+1}^0) &= f(x_s^0 - (x_s^0 - x_{s+1}^0)) \\
&\stackrel{(\text{Lip})}{\leqslant} f(x_s^0) + \langle \nabla f(x_s^0), x_{s+1}^0 - x_s^0 \rangle + \frac{L}{2} \|x_{s+1}^0 - x_s^0\|^2 \\
&= f(x_s^0) - \gamma(n+1) \left\langle \nabla f(x_s^0), \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\rangle \\
&\quad + \frac{\gamma^2(n+1)^2 L}{2} \left\| \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \\
&\stackrel{(\text{Norm})}{=} f(x_s^0) - \frac{\gamma(n+1)}{2} \left[\|\nabla f(x_s^0)\|^2 + \left\| \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \right. \\
&\quad \left. - \left\| \nabla f(x_s^0) - \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \right] \\
&\quad + \frac{\gamma^2(n+1)^2 L}{2} \left\| \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \\
&= f(x_s^0) - \frac{\gamma(n+1)}{2} \left[\|\nabla f(x_s^0)\|^2 - \left\| \nabla f(x_s^0) - \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \right. \\
&\quad \left. - \frac{\gamma(n+1)}{2} \cdot (1 - \gamma(n+1)L) \left\| \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \right].
\end{aligned}$$

Остаётся выбрать такой шаг γ , чтобы выполнялось неравенство

$$\frac{\gamma(n+1)}{2} (1 - \gamma(n+1)L) > 0.$$

Оно соблюдается при условии $\gamma \leqslant \frac{1}{L(n+1)}$, что делает последний член отрицательным и завершает доказательство леммы. \square

Теперь необходимо оценить последний член в результате леммы 11. Для этого доказывается следующая лемма.

Лемма 12 (Лемма 2). Пусть выполняются Предположения 1 и 3. Тогда для алгоритма 2 справедлива следующая оценка:

$$\left\| \nabla f(x_s^0) - \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \leq 2\|\nabla f(x_s^0) - v_s\|^2 + \frac{2L^2}{n+1} \sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2.$$

Доказательство. Будет показано, что

$$\sum_{t=k}^n v_s^t = \frac{1}{n} \sum_{t=k+1}^n (n-t+1) (\nabla f_{\pi_s^t}(x_s^t) - \nabla f_{\pi_s^t}(x_s^{t-1})) + (n-k+1)v_s^k. \quad (8)$$

Докажем это по индукции. При $k = n$ равенство очевидно. Предположим, что утверждение верно для некоторого $\tilde{k} \geq 1$ и покажем, что оно верно для $k = \tilde{k} - 1$:

$$\begin{aligned} \sum_{t=\tilde{k}-1}^n v_s^t &= v_s^{\tilde{k}-1} + \sum_{t=\tilde{k}}^n v_s^t \\ &= v_s^{\tilde{k}-1} + \frac{1}{n} \sum_{t=\tilde{k}+1}^n (n-t+1) (\nabla f_{\pi_s^t}(x_s^t) - \nabla f_{\pi_s^t}(x_s^{t-1})) + (n-\tilde{k}+1)v_s^{\tilde{k}} \\ &\stackrel{(i)}{=} v_s^{\tilde{k}-1} + \frac{1}{n} \sum_{t=\tilde{k}+1}^n (n-t+1) (\nabla f_{\pi_s^t}(x_s^t) - \nabla f_{\pi_s^t}(x_s^{t-1})) \\ &\quad + (n-\tilde{k}+1) \left(v_s^{\tilde{k}-1} + \frac{1}{n} \left(\nabla f_{\pi_s^{\tilde{k}}}(x_s^{\tilde{k}}) - \nabla f_{\pi_s^{\tilde{k}}}(x_s^{\tilde{k}-1}) \right) \right) \\ &= \frac{1}{n} \sum_{t=\tilde{k}}^n (n-t+1) (\nabla f_{\pi_s^t}(x_s^t) - \nabla f_{\pi_s^t}(x_s^{t-1})) + (n-\tilde{k}+2)v_s^{\tilde{k}-1}, \end{aligned}$$

где равенство (i) следует из определения v_s^t в (5) при $\tilde{k} \geq 1$. Индукционный переход доказан, следовательно, равенство (8) выполняется.

Подставляя $k = 0$ в (8) и используя $v_s^0 = v_s$, получаем:

$$\sum_{t=0}^n v_s^t = \frac{1}{n} \sum_{t=1}^n (n-t+1) (\nabla f_{\pi_s^t}(x_s^t) - \nabla f_{\pi_s^t}(x_s^{t-1})) + (n+1)v_s. \quad (9)$$

Оценим интересующий нас член:

$$\begin{aligned}
& \left\| \nabla f(x_s^0) - \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \\
&= \frac{1}{(n+1)^2} \left\| (n+1) \nabla f(x_s^0) - \sum_{t=0}^n v_s^t \right\|^2 \\
&\stackrel{(9)}{=} \frac{1}{(n+1)^2} \left\| (n+1) \nabla f(x_s^0) \right. \\
&\quad \left. - \frac{1}{n} \sum_{t=1}^n (n-t+1) (\nabla f_{\pi_s^t}(x_s^t) - \nabla f_{\pi_s^t}(x_s^{t-1})) \right. \\
&\quad \left. - (n+1)v_s \right\|^2 \\
&\stackrel{(CS)}{\leq} 2 \|\nabla f(x_s^0) - v_s\|^2 \\
&\quad + \frac{2}{(n+1)^2} \left\| \frac{1}{n} \sum_{t=1}^n (n-t+1) (\nabla f_{\pi_s^t}(x_s^t) - \nabla f_{\pi_s^t}(x_s^{t-1})) \right\|^2 \\
&\stackrel{(i)}{\leq} 2 \|\nabla f(x_s^0) - v_s\|^2 \\
&\quad + \frac{2}{(n+1)^2} \left\| \sum_{t=1}^n (\nabla f_{\pi_s^t}(x_s^t) - \nabla f_{\pi_s^t}(x_s^{t-1})) \right\|^2 \\
&\stackrel{(CS)}{\leq} 2 \|\nabla f(x_s^0) - v_s\|^2 + \frac{2}{n+1} \sum_{t=1}^n \|\nabla f_{\pi_s^t}(x_s^t) - \nabla f_{\pi_s^t}(x_s^{t-1})\|^2 \\
&\stackrel{\text{Предп. 1}}{\leq} 2 \|\nabla f(x_s^0) - v_s\|^2 + \frac{2L^2}{n+1} \sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2,
\end{aligned}$$

где неравенство (i) справедливо, поскольку $t \geq 1$ на всём диапазоне суммирования. Полученное неравенство завершает доказательство леммы. \square

Лемма 13 (Лемма 3). *Пусть выполняются Предположения 1 и 3. Если шаг градиентного спуска удовлетворяет ограничению $\gamma \leq \frac{1}{3L}$, то для алгоритма 2 справедлива следующая оценка:*

$$\left\| \nabla f(x_s^0) - \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \leq 9\gamma^2 L^2 \|v_s\|^2 + 36\gamma^2 L^2 n^2 \|v_{s-1}\|^2.$$

Доказательство. Воспользуемся результатом леммы 2, согласно которому:

$$\left\| \nabla f(x_s^0) - \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \leq 2 \|\nabla f(x_s^0) - v_s\|^2 + \frac{2L^2}{n+1} \sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2. \quad (10)$$

В лемме 10 показано, что v_s вычисляется как скользящее среднее стохастических градиентов на предыдущей эпохе:

$$v_s = \tilde{v}_{s-1}^{n+1} = \frac{1}{n} \sum_{t=1}^n \nabla f_{\pi_{s-1}^t}(x_{s-1}^t). \quad (11)$$

Подставляя (11) в (10), получаем:

$$\begin{aligned} \left\| \nabla f(x_s^0) - \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 &\leq 2 \left\| \nabla f(x_s^0) - \frac{1}{n} \sum_{t=1}^n \nabla f_{\pi_{s-1}^t}(x_{s-1}^t) \right\|^2 \\ &\quad + \frac{2L^2}{n+1} \sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2. \end{aligned}$$

Затем, используя соотношение (1),

$$\begin{aligned} &\left\| \nabla f(x_s^0) - \frac{1}{n+1} \sum_{t=0}^n v_s^t \right\|^2 \\ &\leq 2 \left\| \frac{1}{n} \sum_{t=1}^n \left[\nabla f_{\pi_{s-1}^t}(x_s^0) - \nabla f_{\pi_{s-1}^t}(x_{s-1}^t) \right] \right\|^2 \\ &\quad + \frac{2L^2}{n+1} \sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2 \\ &\stackrel{\text{(CS), Предп. 1}}{\leq} \frac{2L^2}{n} \sum_{t=1}^n \|x_{s-1}^t - x_s^0\|^2 + \frac{2L^2}{n+1} \sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2 \\ &\stackrel{\text{(Quad)}}{\leq} \frac{4L^2}{n} \sum_{t=1}^n \|x_{s-1}^t - x_{s-1}^0\|^2 + \frac{4L^2}{n} \sum_{t=1}^n \|x_s^0 - x_{s-1}^0\|^2 \\ &\quad + \frac{2L^2}{n+1} \sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2. \end{aligned} \quad (12)$$

Далее требуется оценить три слагаемых. Начнём с нормы $\sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2$.

$$\sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2 = \gamma^2 \sum_{t=1}^n \|v_s^{t-1}\|^2 = \gamma^2 \sum_{t=0}^{n-1} \|v_s^t\|^2. \quad (13)$$

Далее проводится оценка выражения $\|v_s^t\|^2$. Для $t \geq 1$ справедливо:

$$\begin{aligned} \|v_s^t\|^2 &= \left\| v_s^{t-1} + \frac{1}{n} (\nabla f_{\pi_s^t}(x_s^t) - \nabla f_{\pi_s^t}(x_s^{t-1})) \right\|^2 \\ &\stackrel{\text{(Quad)}}{\leq} \left(1 + \frac{1}{\beta}\right) \|v_s^{t-1}\|^2 + \frac{(1+\beta)L^2}{n^2} \|x_s^t - x_s^{t-1}\|^2 \\ &\stackrel{\text{(Quad)}}{\leq} \left(1 + \frac{1}{\beta}\right)^2 \|v_s^{t-2}\|^2 + \frac{1}{n^2} \left(1 + \frac{1}{\beta}\right) (1+\beta)L^2 \|x_s^{t-1} - x_s^{t-2}\|^2 \\ &\quad + \frac{1}{n^2} (1+\beta)L^2 \|x_s^t - x_s^{t-1}\|^2 \\ &\stackrel{\text{(Quad)}}{\leq} \left(1 + \frac{1}{\beta}\right)^t \|v_s\|^2 + \frac{1}{n^2} (1+\beta)L^2 \sum_{k=1}^t \left(1 + \frac{1}{\beta}\right)^{k-1} \|x_s^{t-k+1} - x_s^{t-k}\|^2 \\ &\stackrel{\text{(Quad)}}{\leq} \beta=t \left(1 + \frac{1}{t}\right)^t \|v_s\|^2 + \frac{1}{n^2} (1+t) \left(1 + \frac{1}{t}\right)^t L^2 \sum_{k=1}^t \|x_s^k - x_s^{k-1}\|^2. \end{aligned}$$

Далее, используя свойство показательной функции $\left(\left(1 + \frac{1}{t}\right)^t \leq e\right)$ и факт, что $t \leq n-1$, а также неравенство (13), получаем важную оценку (для $0 \leq t \leq n-1$, поскольку при $t=0$ выполняется $\|v_s^t\|^2 = \|v_s\|^2$, и искомое неравенство становится тривиальным):

$$\|v_s^t\|^2 \leq e\|v_s\|^2 + \frac{eL^2}{n} \sum_{k=1}^t \|x_s^k - x_s^{k-1}\|^2. \quad (14)$$

Теперь подставим выражение из (14) в (13) и получим:

$$\begin{aligned} \sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2 &= \gamma^2 \sum_{t=0}^{n-1} \|v_s^t\|^2 \leq e\gamma^2 n \|v_s\|^2 + \frac{e\gamma^2 L^2}{n} \sum_{t=0}^{n-1} \sum_{k=1}^t \|x_s^k - x_s^{k-1}\|^2 \\ &\leq e\gamma^2 n \|v_s\|^2 + e\gamma^2 L^2 \sum_{t=1}^{n-1} \|x_s^t - x_s^{t-1}\|^2 \\ &\leq e\gamma^2 n \|v_s\|^2 + e\gamma^2 L^2 \sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2. \end{aligned}$$

Непосредственно раскрывая сумму $\sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2$, получаем требуемую оценку:

$$\sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2 \leq \frac{e\gamma^2 n \|v_s\|^2}{1 - e\gamma^2 L^2} \stackrel{e < 3}{\leq} \frac{3\gamma^2 n \|v_s\|^2}{1 - 3\gamma^2 L^2}.$$

Для завершения данной части доказательства остаётся выбрать подходящее значение γ . В лемме 11 требуется выполнение условия $\gamma \leq \frac{1}{L(n+1)}$. Там допускаются даже большие значения γ . Проведём оценку полученного выражения при $\gamma \leq \frac{1}{L(n+1)} \leq \frac{1}{3L}$. Ниже приводится окончательная оценка соответствующей нормы:

$$\sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2 \leq \frac{9}{2} \gamma^2 n \|v_s\|^2. \quad (15)$$

Продолжим оценивание неравенства (12), рассмотрев слагаемое $\sum_{t=1}^n \|x_{s-1}^t - x_{s-1}^0\|^2$.

$$\begin{aligned} \sum_{t=1}^n \|x_{s-1}^t - x_{s-1}^0\|^2 &= \gamma^2 \sum_{t=1}^n \left\| \sum_{k=0}^{t-1} v_{s-1}^k \right\|^2 \\ &\stackrel{(CS)}{\leq} \gamma^2 \sum_{t=1}^n t \sum_{k=0}^{t-1} \|v_{s-1}^k\|^2 \leq \gamma^2 n^2 \sum_{t=0}^{n-1} \|v_{s-1}^t\|^2. \end{aligned} \quad (16)$$

Отметим, что ранее уже была получена оценка для $\|v_s^t\|^2$ при $0 \leq t \leq n-1$, см. (14). Аналогичную оценку можно применить и к соответствующим величинам на эпохе $(s-1)$, после чего записывается:

$$\|v_{s-1}^t\|^2 \leq e\|v_{s-1}\|^2 + \frac{eL^2}{n} \sum_{k=1}^t \|x_{s-1}^k - x_{s-1}^{k-1}\|^2. \quad (17)$$

Теперь подставим выражение из (17) в (16), чтобы получить:

$$\begin{aligned}
\sum_{t=1}^n \|x_{s-1}^t - x_{s-1}^0\|^2 &\leq \gamma^2 n^2 \sum_{t=0}^{n-1} \left(e\|v_{s-1}\|^2 + \frac{eL^2}{n} \sum_{k=1}^t \|x_{s-1}^k - x_{s-1}^{k-1}\|^2 \right) \\
&\leq \gamma^2 n^3 e\|v_{s-1}\|^2 + e\gamma^2 L^2 n \sum_{t=0}^{n-1} \sum_{k=1}^t \|x_{s-1}^k - x_{s-1}^{k-1}\|^2 \\
&\leq \gamma^2 n^3 e\|v_{s-1}\|^2 + e\gamma^2 L^2 n^2 \sum_{t=1}^{n-1} \|x_{s-1}^t - x_{s-1}^{t-1}\|^2 \\
&\leq \gamma^2 n^3 e\|v_{s-1}\|^2 + e\gamma^2 L^2 n^2 \sum_{t=1}^n \|x_{s-1}^t - x_{s-1}^{t-1}\|^2. \quad (18)
\end{aligned}$$

Заметим, что слагаемое $\sum_{t=1}^n \|x_s^t - x_s^{t-1}\|^2$ уже было оценено ранее, см. (15). Аналогичную оценку можно провести и для соответствующего слагаемого на эпохе $(s-1)$, после чего записывается:

$$\sum_{t=1}^n \|x_{s-1}^t - x_{s-1}^{t-1}\|^2 \leq \frac{9}{2} \gamma^2 n \|v_{s-1}\|^2. \quad (19)$$

Подставляя выражение из (19) в (18), получаем:

$$\sum_{t=1}^n \|x_{s-1}^t - x_{s-1}^0\|^2 \leq 3\gamma^2 n^3 \|v_{s-1}\|^2 + \frac{27}{2} \gamma^4 L^2 n^3 \|v_{s-1}\|^2.$$

Используя выбранное ограничение $\gamma \leq \frac{1}{3L}$,

$$\sum_{t=1}^n \|x_{s-1}^t - x_{s-1}^0\|^2 \leq 3\gamma^2 n^3 \|v_{s-1}\|^2 + \frac{3}{2} \gamma^2 n^3 \|v_{s-1}\|^2 = \frac{9}{2} \gamma^2 n^3 \|v_{s-1}\|^2. \quad (20)$$

Остаётся оценить слагаемое $\sum_{t=1}^n \|x_s^0 - x_{s-1}^0\|^2$. Получаемая оценка аналогична предыдущей:

$$\begin{aligned}
\sum_{t=1}^n \|x_s^0 - x_{s-1}^0\|^2 &= \gamma^2 \sum_{t=1}^n \left\| \sum_{k=0}^{n-1} v_{s-1}^{k-1} \right\|^2 \\
&\stackrel{\text{(Quad)}}{\leq} \gamma^2 \sum_{t=1}^n n \sum_{k=0}^{n-1} \|v_{s-1}^k\|^2 \leq \gamma^2 n^2 \sum_{t=0}^{n-1} \|v_{s-1}^t\|^2.
\end{aligned}$$

Получается оценка, аналогичная (16). Действуя аналогично предыдущему случаю, получаем:

$$\sum_{t=1}^n \|x_s^0 - x_{s-1}^0\|^2 \leq \frac{9}{2}\gamma^2 n^3 \|v_{s-1}\|^2. \quad (21)$$

Теперь можно подставить верхние оценки из (15), (20), (21) в (12) и получить:

$$\begin{aligned} \left\| \nabla f(\omega_s) - \frac{1}{n+1} \sum_{i=0}^n v_s^i \right\|^2 &\leq 18\gamma^2 L^2 n^2 \|v_{s-1}\|^2 + 18\gamma^2 L^2 n^2 \|v_{s-1}\|^2 + 9\gamma^2 L^2 \|v_s\|^2 \\ &= 9\gamma^2 L^2 \|v_s\|^2 + 36\gamma^2 L^2 n^2 \|v_{s-1}\|^2, \end{aligned}$$

что завершает доказательство. \square

Теорема 3 (Теорема 1). Пусть выполнены Предположения 1 и 3. Тогда для достижения ε -точности, где $\varepsilon^2 = \frac{1}{S} \sum_{s=1}^S \|\nabla f(x_s^0)\|^2$, алгоритму 2 при шаге $\gamma \leq \frac{1}{20L(n+1)}$ требуется

$$\mathcal{O}\left(\frac{nL}{\varepsilon^2}\right) \text{ итераций и вызовов оракула.}$$

Доказательство. Объединяя результат леммы 11 с результатом леммы 13, получаем:

$$\begin{aligned} f(x_{s+1}^0) &\leq f(x_s^0) - \frac{\gamma(n+1)}{2} \|\nabla f(x_s^0)\|^2 \\ &\quad + \frac{\gamma(n+1)}{2} (9\gamma^2 L^2 \|v_s\|^2 + 36\gamma^2 L^2 n^2 \|v_{s-1}\|^2). \end{aligned}$$

Вычтем $f(x^*)$ из обеих частей неравенства:

$$\begin{aligned} f(x_{s+1}^0) - f(x^*) &\leq f(x_s^0) - f(x^*) - \frac{\gamma(n+1)}{2} \|\nabla f(x_s^0)\|^2 \\ &\quad + \frac{\gamma(n+1)}{2} (9\gamma^2 L^2 \|v_s\|^2 + 36\gamma^2 L^2 n^2 \|v_{s-1}\|^2) \\ &= f(x_s^0) - f(x^*) - \frac{\gamma(n+1)}{4} \|\nabla f(x_s^0)\|^2 \\ &\quad + \frac{\gamma(n+1)}{2} (9\gamma^2 L^2 \|v_s\|^2 + 36\gamma^2 L^2 n^2 \|v_{s-1}\|^2) \\ &\quad - \frac{\gamma(n+1)}{4} \|\nabla f(x_s^0)\|^2. \end{aligned}$$

Затем, преобразуя последнее слагаемое с использованием неравенства (Quad) при $\beta = 1$, получаем:

$$\begin{aligned} f(x_{s+1}^0) - f(x^*) &\leq f(x_s^0) - f(x^*) - \frac{\gamma(n+1)}{4} \|\nabla f(x_s^0)\|^2 \\ &\quad + \frac{\gamma(n+1)}{2} (9\gamma^2 L^2 \|v_s\|^2 + 36\gamma^2 L^2 n^2 \|v_{s-1}\|^2) \\ &\quad - \frac{\gamma(n+1)}{8} \|v_s\|^2 + \frac{\gamma(n+1)}{4} \|v_s - \nabla f(x_s^0)\|^2. \end{aligned}$$

Применяя лемму 13 к выражению $\|v_s - \nabla f(x_s^0)\|^2$ (в частности, используя оценки $\frac{4L^2}{n} \cdot (20)$ и $\frac{4L^2}{n} \cdot (21)$),

$$\begin{aligned} f(x_{s+1}^0) - f(x^*) &\leq f(x_s^0) - f(x^*) - \frac{\gamma(n+1)}{4} \|\nabla f(x_s^0)\|^2 \\ &\quad + \frac{\gamma(n+1)}{2} (9\gamma^2 L^2 \|v_s\|^2 + 36\gamma^2 L^2 n^2 \|v_{s-1}\|^2) \\ &\quad - \frac{\gamma(n+1)}{8} \|v_s\|^2 + \frac{\gamma(n+1)}{4} \cdot 36\gamma^2 L^2 n^2 \|v_{s-1}\|^2. \end{aligned}$$

Объединяя подобные слагаемые,

$$\begin{aligned} f(x_{s+1}^0) - f(x^*) &+ \frac{\gamma(n+1)}{4} \|\nabla f(x_s^0)\|^2 \\ &\leq f(x_s^0) - f(x^*) - \frac{\gamma(n+1)}{8} (1 - 36\gamma^2 L^2) \|v_s\|^2 \\ &\quad + \gamma(n+1) \cdot 27\gamma^2 L^2 n^2 \|v_{s-1}\|^2. \end{aligned} \tag{22}$$

Используя условие $\gamma \leq \frac{1}{20L(n+1)}$ (заметим, что это наименьший из всех ранее использованных шагов, поэтому все предыдущие переходы остаются корректными), получаем:

$$\begin{aligned} f(x_{s+1}^0) - f(x^*) &+ \frac{1}{10} \gamma(n+1) \|v_s\|^2 + \frac{\gamma(n+1)}{4} \|\nabla f(\omega_s)\|^2 \\ &\leq f(x_s^0) - f(x^*) + \frac{1}{10} \gamma(n+1) \|v_{s-1}\|^2. \end{aligned}$$

Обозначив $\Delta_s = f(x_{s+1}^0) - f(x^*) + \frac{1}{10} \gamma(n+1) \|v_s\|^2$, получаем:

$$\frac{1}{S} \sum_{s=1}^S \|\nabla f(x_s^0)\|^2 \leq \frac{4 [\Delta_0 - \Delta_S]}{\gamma(n+1) S}.$$

В качестве критерия выбирается $\varepsilon^2 = \frac{1}{S} \sum_{s=1}^S \|\nabla f(x_s^0)\|^2$. Таким образом, для достижения ε -точности требуется $\mathcal{O}\left(\frac{L}{\varepsilon^2}\right)$ эпох и $\mathcal{O}\left(\frac{nL}{\varepsilon^2}\right)$ итераций. Кроме того, отмечается, что сложность по оракулу для рассматриваемого алгоритма также составляет $\mathcal{O}\left(\frac{nL}{\varepsilon^2}\right)$, поскольку на каждой итерации вычисляется стохастический градиент в двух точках. Это завершает доказательство. \square

B.3 Случай сильной выпуклости

Теорема 4 (Теорема 2). *Пусть выполняются Предположения 1 и 2. Тогда алгоритм 2 при шаге $\gamma \leq \frac{1}{20L(n+1)}$ достигает ε -точности по функционалу, то есть*

$$\varepsilon = f(x_{S+1}^0) - f(x^*),$$

за

$$\mathcal{O}\left(\frac{nL}{\mu} \log \frac{1}{\varepsilon}\right)$$

итераций и вызовов стохастического оракула.

Доказательство. Согласно Предположению 2, функция f является сильно выпуклой, а значит, условие Поляка-Лоясевича (PL) выполняется автоматически. Тогда

$$\begin{aligned} & f(x_{s+1}^0) - f(x^*) + \frac{\gamma\mu(n+1)}{2} (f(x_s^0) - f(x^*)) \\ & \leq f(x_{s+1}^0) - f(x^*) + \frac{\gamma(n+1)}{4} \|\nabla f(x_s^0)\|^2. \end{aligned}$$

Используя неравенство (22), получаем:

$$\begin{aligned} & f(x_{s+1}^0) - f(x^*) + \frac{\gamma\mu(n+1)}{2} (f(x_s^0) - f(x^*)) \leq f(x_s^0) - f(x^*) \\ & \quad - \frac{\gamma(n+1)}{8} (1 - 36\gamma^2 L^2) \|v_s\|^2 \\ & \quad + 27\gamma^3(n+1)L^2 n^2 \|v_{s-1}\|^2. \end{aligned}$$

При $\gamma \leq \frac{1}{20L(n+1)}$ и $n \geq 2$, имеем:

$$\begin{aligned} f(x_{s+1}^0) - f(x^*) + \frac{1}{10}\gamma(n+1)\|v_s\|^2 &\leq \left(1 - \frac{\gamma\mu(n+1)}{2}\right)(f(x_s^0) - f(x^*)) \\ &+ \frac{1}{10}\gamma(n+1)\left(1 - \frac{\gamma\mu(n+1)}{2}\right)\|v_{s-1}\|^2. \end{aligned}$$

Обозначив

$$\Delta_s = f(x_{s+1}^0) - f(x^*) + \frac{1}{10}\gamma(n+1)\|v_s\|^2,$$

получаем рекурсивное неравенство:

$$\Delta_{s+1} \leq \left(1 - \frac{\gamma\mu(n+1)}{2}\right)\Delta_s.$$

Переходя к рекурсии по всем эпохам:

$$f(x_{S+1}^0) - f(x^*) \leq \Delta_S \leq \left(1 - \frac{\gamma\mu(n+1)}{2}\right)^{S+1} \Delta_0.$$

Полагая $\varepsilon = f(x_{S+1}^0) - f(x^*)$, заключаем, что для достижения ε -точности достаточно

$$\mathcal{O}\left(\frac{L}{\mu} \log \frac{1}{\varepsilon}\right)$$

эпох и

$$\mathcal{O}\left(\frac{nL}{\mu} \log \frac{1}{\varepsilon}\right)$$

итераций.

Кроме того, сложность по числу вызовов стохастического градиентного оракула также составляет

$$\mathcal{O}\left(\frac{nL}{\mu} \log \frac{1}{\varepsilon}\right),$$

так как на каждой итерации вычисляется градиент в не более чем двух точках. Тем самым доказательство завершено. \square