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

БЕЗГРАДИЕНТНЫЕ АЛГОРИТМЫ ДЛЯ РЕШЕНИЯ СТОХАСТИЧЕСКИХ ЗАДАЧ ВЫПУКЛОЙ ОПТИМИЗАЦИИ С БЕСКОНЕЧНЫМ ШУМОМ

(бакалаврская работа)

Студент:
Корнилов Никита Максимович

____
(подпись студента)

Научный руководитель:
Гасников Александр Владимирович,
д-р физ.-мат. наук

____
(подпись научного руководителя)

Консультант *(при наличии)*:
____
____
(подпись консультанта)

Москва 2023

Аннотация

Во многих задачах оптимизации информация о градиенте целевой функции недоступна, и доступ к значениям функции можно получить только через оракул по типу черного ящика. Такие условия требуют использования методов нулевого порядка. В данной работе изучается задача негладкой оптимизации на выпуклом компакте со стохастическим шумом с тяжелыми хвостами, т.е. шум с $(1+\kappa)$-м ограниченным моментом, и с враждебным шумом в значениях функции. Мы предлагаем два новых алгоритма, оптимальных с точки зрения количества вызовов двухточечного оракула. Их оракульная сложность пропорциональна $\left(\sqrt{d}/\varepsilon\right)^{\frac{1+\kappa}{\kappa}}$ с высокой вероятностью и по математическому ожиданию, где d – размерность пространства, а ε – точность по значению функции.

Наши алгоритмы построены на эффективных методах первого порядка. Первый алгоритм основан на устойчивом к шуму с тяжелыми хвостами зеркальном спуске. Второй алгоритм основан на методе клиппирования градиента. В работе обсуждаются различия между этими двумя алгоритмами, а также рекомендации по настройке параметров. Кроме того, для целевых функций, удовлетворяющих условию острого минимума, предлагается более быстрый алгоритм с использованием техники рестартов. Особое внимание уделяется вопросу о том, насколько большим может быть враждебный шум, чтобы гарантировать оптимальность алгоритмов.

Оглавление

Глава 1

Введение

Обозначения. Для $p \in [1,2]$, мы используем l_p-норму, т.е.

$$\|x\|_p = \left(\sum_{k=1}^{d} |x_k|^p \right)^{1/p}.$$

Соответствующая ей сопряженная норма обозначается

$$\|y\|_q = \max_x \{ \langle x, y \rangle | \; \|x\|_p \le 1 \},$$

где q определяется равенством $\frac{1}{q} + \frac{1}{p} = 1$.

Мы используем $\langle x,y \rangle = \sum_{k=1}^{d} x_k y_k$ для скалярного произведения между $x,y \in \mathbb{R}^d$. А $B^p = \{x \in \mathbb{R}^d \mid \|x\|_p \le 1\}$ и $S^p = \{x \in \mathbb{R}^d \mid \|x\|_p = 1\}$ обозначают единичный ℓ_p-шар и единичную ℓ_p-сферу соответственно.

Полное математическое ожидание случайной величины X записывается через $\mathbb{E}[X]$. А мат ожидание по случайным величинам $Y_1, \ldots, Y_n$ – через $\mathbb{E}_{Y_1,\ldots,Y_n}[X]$. Условное мат ожидание при условии случайных величин $x_k, \ldots, x_1$ обозначается как $\mathbb{E}[\cdot | x_k, \ldots, x_1] \overset{\text{def}}{=} \mathbb{E}_{|\le k}[\cdot]$ для краткости.

Постановка задачи. Мы рассматриваем негладкую задачу выпуклой стохастической минимизации на выпуклом компакте $\mathcal{X} \subset \mathbb{R}^d$:

$$\min_{x \in \mathcal{X}} \left\{ f(x) \triangleq \mathbb{E}_\xi [f(x,\xi)] \right\}, \tag{1.1}$$

где значения целевой функции доступны только через зашумленный оракул нулевого порядка

$$\phi(x,\xi) = f(x,\xi) + \delta(x). \tag{1.2}$$

Значения оракула можно получить через процедуру по типу черного ящика, например симуляцию. Мы рассматриваем только двухточечный оракул, это означает, что для двух заданных точек $x,y \in \mathcal{X}$ мы получаем $\phi(x,\xi)$ и $\phi(y,\xi)$ с одним и тем же ξ. Функция $\phi(x,\xi)$ может рассматриваться как зашумленное приближение Липшецевой функции $f(x,\xi)$. Эта зашумленность может быть детерминированной, стохастической или враждебной.

Актуальность. Методы нулевого порядка изучались в большом количестве работ, см., напр. [1, 2] и ссылки в них. В частности, при различных предположениях об оракуле черного ящика (в зашумленной или бесшумной установке) была получена оптимальная сложность оракула [3, 4, 5, 6, 7, 8]. Такая постановка может быть полезна при поиске гиперпараметров, решении задач с многорукими бандитами, а также в ситуациях, когда подсчёт значения целевой функции стоит дорого, например, в задачах Reinforcement Learning. В работе [9] был получен алгоритм, в которой допускался шум с ограниченной дисперсией. Многие результаты из [9] обобщаются в Главе 2. Однако до сих пор не было предложено метода нулевого порядка устойчивого к шуму с тяжелыми хвостами.

В этой работе мы пользуемся техникой клиппированния. Эта техника становится все более популярной. Она позволяет бороться с тяжелыми хвостами при обучении нейронных сетей, задач машинного обучения и стохастической оптимизации, делая процедуру устойчивее. Её также используют для получения гарантий сходимости с высокой вероятностью [10, 11, 12].

Помимо этого, следуя работе [13], в которой был предложен устойчивый к тяжелому шуму зеркальный спуск первого порядка, мы обобщаем его на нулевой порядок.

Предположения. Для выпуклого множества $\mathcal{X} \subset \mathbb{R}^d$ и константы $\tau > 0$, мы вводим обозначение $\mathcal{X}_\tau = \mathcal{X} + \tau B^2$.

Предположение 1 (Выпуклость). *Существует такая константа $\tau > 0$, что функция $f(x, \xi)$ выпукла по x на $\mathcal{X}_\tau$ для любого ξ.*

Из этого Предположения следует, что $f(x)$ также выпукла на $\mathcal{X}$.

Предположение 2 (Липшецевость). *Существует такая константа $\tau > 0$, что функция $f(x, \xi)$ $M_2(\xi)$-Липшецева по x в l_2-норме т.е. для любых $x_1, x_2 \in \mathcal{X}_\tau$*

$$|f(x_1, \xi) - f(x_2, \xi)| \leq M_2(\xi)\|x_1 - x_2\|_2.$$

К тому же, существует $\kappa \in (0{,}1]$ и M_2 такие, что $\mathbb{E}_\xi[M_2(\xi)^{1+\kappa}] \leq M_2^{1+\kappa}$.

Предположение 3 (Ограниченность враждебного шума). *Для любых $x \in \mathcal{X} : |\delta(x)| \leq \Delta < \infty$.*

Глава 2

Безградиентная оптимизация и Зеркальный Спуск

2.0.1 Безградиентная оптимизация

В этом главе мы представим основные понятия и обозначения, которые используются для построения безградиентных методов.

Мы рассматриваем равномерное семплирование с единичной евклидовой сферы

$$\mathbf{e} \sim \text{Uniform}(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\}) \overset{\text{def}}{=} U(S^2).$$

Мы также определяем следующую гладкую аппроксимацию целевой функции $f(x)$

$$\hat{f}_\tau(x) \triangleq \mathbb{E}_{\mathbf{e}}[f(x + \tau\mathbf{e})], \tag{2.1}$$

где $\mathbf{e} \sim U(S^2)$ и $\tau > 0$.

Следующая Лемма показывает качество данной аппроксимации.

Лемма 2.0.1. *Пусть Предположения 1,2 выполнены. Тогда*

1. *Функция $\hat{f}_\tau(x)$ выпукла, M_2-Липшецева на $\mathcal{X}$ и удовлетворяет неравенству*
$$\sup_{x\in\mathcal{X}} |\hat{f}_\tau(x) - f(x)| \leq \tau M_2. \tag{2.2}$$

2. *Функция $\hat{f}_\tau(x)$ дифференцируема на $\mathcal{X}$ со следующим градиентом*
$$\nabla \hat{f}_\tau(x) = \mathbb{E}_{\mathbf{e}}\left[\frac{d}{\tau} f(x + \tau\mathbf{e})\mathbf{e}\right]. \tag{2.3}$$

Доказательство этой леммы может быть найдено в [14, Теорема 2.1].

Следуя работе [7], рассмотрим следующий случайный вектор, который строится на двух точках и соответствующих двух значениях оракула нулевого порядка

$$\begin{aligned} g(x, \xi, \mathbf{e}) &= \frac{d}{2\tau}(\phi(x + \tau\mathbf{e}, \xi) - \phi(x - \tau\mathbf{e}, \xi))\mathbf{e} \\ &= \frac{d}{2\tau}(f(x + \tau\mathbf{e}, \xi) + \delta(x + \tau\mathbf{e}) \\ &- (f(x - \tau\mathbf{e}, \xi) + \delta(x - \tau\mathbf{e})))\mathbf{e}. \end{aligned} \tag{2.4}$$

Этот вектор будет несмещенной оценкой (если убрать враждебный шум) градиента $\hat{f}_\tau(x)$. Интуитивно это можно понять по аналогии с разностной схемой оценки градиента по двум точкам. Также у этого вектора ограниченный $(1 + \kappa)$-ый момент, подробности приведены в лемме ниже.

Лемма 2.0.2. *При предположениях 1, 2 и 3, для $q \in [2, +\infty)$, верно следующее неравенство*

$$\mathbb{E}_{\xi,\mathbf{e}}[\|g(x,\xi,\mathbf{e})\|_q^{1+\kappa}] \leq 2^\kappa \left(\frac{\sqrt{d}}{2^{1/4}} a_q M_2\right)^{1+\kappa} + 2^\kappa \left(\frac{d a_q \Delta}{\tau}\right)^{1+\kappa} = \sigma_q^{1+\kappa},$$

где $a_q \overset{def}{=} d^{\frac{1}{q}-\frac{1}{2}} \min\{\sqrt{32\ln d - 8}, \sqrt{2q-1}\}$.

Все алгоритмы ниже будут пытаться минимизировать именно сглаженную аппроксимацию функции $f(x)$, при этом полученное решение хорошей точности будет подходить и для функции $f(x)$ при малом τ.

2.0.2 Устойчивый зеркальный спуск

В этой главе мы рассмотрим стандартный зеркальный спуск из работы [15] и обобщим его для равномерно выпуклых прокс-функций, также как это сделано в работе [13]. Мы также представим все необходимые результаты о сходимости алгоритма.

Определение 2.0.1. *Пусть даны дифференцируемая выпуклая функция* $\psi : \mathbb{R}^d \to \mathbb{R}$ *, степень* $r \geq 2$ *и константа* $K > 0$. ψ *называется* (K,r)*-равномерно выпуклой по* ℓ_p*-норме, если для любого* $x,y \in \mathbb{R}^d$,

$$\psi(y) - \psi(x) - \langle \nabla\psi(x), y - x\rangle \geq \frac{K}{r}\|x - y\|_p^r. \tag{2.5}$$

Когда $r = 2$ определение (K,r)-равномерной выпуклости совпадает с определением K-сильной выпуклости. Примеры таких функций при $r > 2$ могут быть получены из следующей леммы.

Лемма 2.0.3. *Для* $\kappa \in (0,1], q \in [1 + \kappa, \infty)$ *и* p *такого, что* $\frac{1}{q} + \frac{1}{p} = 1$, *рассмотрим*

$$K_q \overset{def}{=} 10 \max\left\{1, (q-1)^{\frac{1+\kappa}{2}}\right\}. \tag{2.6}$$

Тогда

$$\phi_p(x) \overset{def}{=} \frac{\kappa}{1+\kappa}\|x\|_p^{\frac{1+\kappa}{\kappa}} \tag{2.7}$$

$\left(K_q^{-\frac{1}{\kappa}}, \frac{1+\kappa}{\kappa}\right)$*-равномерно выпуклая по* ℓ_p*-норме.*

Стохастический Зеркальный Спуск или Stochastic Mirror Descent (SMD) обобщает Стохастический Градиентный Спуск на задачи оптимизации на различных множествах $\mathcal{X}$, позволяя при удачном выборе параметров уменьшить константы в верхних границах сходимости или упростить задаччу проектирования на множество.

Теперь вкратце опишем его. Пусть функция $\Psi : \mathbb{R}^d \to \mathbb{R}$ является (K,r)-равномерно выпуклой по ℓ_p-норме.

Её сопряженная функция и дивергенция Брегмана определяются соответственно как

$$\Psi^*(y) = \sup_{x\in\mathbb{R}^d}\{\langle x,y\rangle - \Psi(x)\} \quad \text{и} \quad D_\Psi(y,x) = \Psi(y) - \Psi(x) - \langle \nabla\Psi(x), y - x\rangle.$$

Шаг Stochastic Mirror Descent с размером шага ν и вектором обновления g_{k+1} задаётся по формулам:

$$y_{k+1} = \nabla(\Psi^*)(\nabla\Psi(x_k) - \nu g_{k+1}), \quad x_{k+1} = \arg\min_{x\in\mathcal{X}} D_\Psi(x, y_{k+1}). \tag{2.8}$$

С помощью предположений на Ψ можно доказать, что шаги корректно определены, а $(\nabla\Psi)^{-1} = \nabla\Psi^*$. Отображение $\nabla\Psi$ называется зеркальным отображением. Суть алгоритма заключается в следующем:

1. $\nabla\Psi$ переводит точку x в двойственное пространство $\mathcal{X}^*$,
2. В этом пространстве происходит шаг градиентного спуска с вектором g_{k+1},
3. $(\nabla\Psi)^{-1} = \nabla\Psi^*$ возвращает полученный вектор обратно в пространство $\mathcal{X}$,
4. $D_\Psi(\cdot,\cdot)$ служит заменой Евклидовой метрике, тем самым на последнем шаге происходит "проекция"на множество $\mathcal{X}$.

Для SMD (2.8) со стандартной 1-сильно выпуклой прокс-функцией Ψ, теория сходимости хорошо изучена, например, в лекциях [16]. Следующая теорема обобщает эти результаты на равномерно выпуклые функции Ψ, где $x^* = \arg\min\limits_{x\in\mathcal{X}} f(x)$ обозначает решение задачи (1.1).

Теорема 2.0.4. *Пусть $\kappa \in (0,1], p \in [1,\infty]$ и прокс-фукнция Ψ_p, которая является $\left(1, \frac{1+\kappa}{\kappa}\right)$-равномерно выпуклой по ℓ_p-норме заданы. Тогда для SMD, описанного в (2.8), через T итерация с любыми векторами $g_k \in \mathbb{R}^d, k \in \overline{1,T}$ и начальной точкой $x_0 = \arg\min\limits_{x\in\mathcal{X}} \Psi_p(x)$ выполнено*

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle g_{k+1}, x_k - x^*\rangle \leq \frac{\kappa}{\kappa+1}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\|g_{k+1}\|_q^{1+\kappa}, \tag{2.9}$$

где $R_0^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}D_{\Psi_p}(x^*, x_0)$ расстояние между начальной точкой x_0 и решением x^*.

Доказательство этой теоремы можно найти в [13, Теорема 6]. Заметим, что при $\kappa = 1$, Ψ_p является 1-сильно выпуклой функцией.

Глава 3

Безградиентный Алгоритм с Устойчивым SMD

3.0.1 Алгоритм и Теорема Сходимости

Основная идея предлагаемого алгоритма Zeroth-Order Robust SMD состоит в том, чтобы объединить вышеупомянутый алгоритм Robust SMD (2.8) с аппроксимацией двухточечного градиента (2.4). Первый позволяет работать с тяжелыми хвостами распределения градиентной аппроксимации, а второй позволяет справиться с негладкостью целевой функции в (1.1).

Следующая теорема предоставляет оценки на скорость сходимости алгоритма 1, а также оптимальные для этого параметры.

Теорема 3.0.1. *Пусть функция f, удовлетворяющая Предположениям 1, 2, 3, $q \in [1+\kappa, \infty]$, количество итераций T, константа сглаживания $\tau > 0$ заданы заранее. Выберем $\left(1, \frac{1+\kappa}{\kappa}\right)$-равномерно выпуклую по ℓ_p-норме прокс-функцию $\Psi_p(x)$ (К примеру, $\Psi_p(x) = K_q^{1/\kappa}\phi_p(x)$, где K_q, ϕ_p определены в (2.6) и (2.7) соответственно). Установим размер шага равный $\nu = \frac{R_0^{1/\kappa}}{\sigma_q} T^{-\frac{1}{1+\kappa}}$ с σ_q из Леммы 2.0.2, расстоянием от начальной точки x_0 до решения x^* $R_0^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa} D_{\Psi_p}(x^*, x_0)$ и диаметром компакта $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=}$*

Algorithm 1 Zeroth-Order Robust SMD Algorithm

procedure ZERO ROBUST SMD(Количество итераций T, размер шага ν, прокс-функция Ψ_p, константа сглавживания τ)
  $x_0 \leftarrow \arg\min\limits_{x \in \mathcal{X}} \Psi_p(x)$
  for $k = 0, 1, \ldots, T-1$ **do**
    Независимо отсемплировать $\mathbf{e}_k \sim \mathrm{Uniform}(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\})$
    Независимо отсемплировать ξ_k
    Вычислить $g_{k+1} = \frac{d}{2\tau}(\phi(x_k + \tau\mathbf{e}_k, \xi_k) - \phi(x_k - \tau\mathbf{e}_k, \xi_k))\mathbf{e}_k$
    Вычислить $y_{k+1} \leftarrow \nabla(\Psi_p^*)(\nabla\Psi_p(x_k) - \nu g_{k+1})$
    Вычислить $x_{k+1} \leftarrow \arg\min\limits_{x \in \mathcal{X}} D_{\Psi_p}(x, y_{k+1})$
  end for
  return $\overline{x}_T \leftarrow \frac{1}{T}\sum\limits_{k=0}^{T-1} x_k$
end procedure

$\frac{1+\kappa}{\kappa} \sup\limits_{x,y \in \mathcal{X}} D_{\Psi_p}(x,y)$. *Пусть* $\overline{x}_T$ *результат работы Алгоритма 1 с заданными выше параметрами.*

1. *Тогда имеем следующую оценку*

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_0\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \tag{3.1}$$

где $\sigma_q^{1+\kappa} = 2^\kappa \left(\frac{\sqrt{d}}{2^{1/4}} a_q M_2\right)^{1+\kappa} + 2^\kappa \left(\frac{d a_q \Delta}{\tau}\right)^{1+\kappa}$.

2. *К тому же, с оптимальным* $\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4R_0 d a_q \Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$ *имеем*

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_0 d a_q\Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{2\sqrt{d}a_q M_2 R_0}{T^{\frac{\kappa}{1+\kappa}}}. \tag{3.2}$$

Набросок доказательства 3.0.1. Доказательство, основано на Теореме 2.0.4

и неравенстве (2.9), которые дают оценку

$$\underbrace{\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\langle g_{k+1}, x_k - x^*\rangle\right]}_{1} \leq \underbrace{\mathbb{E}\left[\frac{\kappa}{\kappa+1}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T}\right]}_{2} + \underbrace{\mathbb{E}\left[\frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\|g_{k+1}\|_q^{1+\kappa}\right]}_{3}. \tag{3.3}$$

1 слагаемое в (3.3) из-за выпуклости и аппроксимационных свойств $\hat{f}_\tau(x)$ из Леммы 2.0.1, а также из-за неравенства концентрации из Леммы 6.1.6 можно ограничить сверху

$$1 \geq \mathbb{E}[f(\overline{x}_T)] - f(x^*) - 2M_2\tau - \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi.$$

3 слагаемое в (3.3) можно ограничить из-за ограниченности $(1+\kappa)$-го момента, полученного в Лемме 2.0.2, как

$$3 \leq \frac{\nu^\kappa}{1+\kappa}\sigma_q^{1+\kappa}.$$

Объединяя эти оценки сверху, мы получим

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\sigma_q^{1+\kappa}.$$

Осталось только выбрать оптимальный размер шага $\nu = \frac{R_0^{1/\kappa}}{\sigma_q}T^{-\frac{1}{1+\kappa}}, \tau$ и закончить доказательство. □

Для полного доказательства мы отсылаем читателя в Главу 6.2.

3.0.2 Обсуждение

Максимальный уровень допустимого враждебного шума. Пусть $\varepsilon > 0$ — желаемая точность с точки зрения значения функции, т.е. наша цель — гарантировать $\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq \varepsilon$. Согласно теореме 3.0.1 при отсутствии враждебного шума, т.е. при $\Delta = 0$, количество итераций для достижения точности ε составляет $T = \left(\frac{R_0\sqrt{d}a_qM_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}$, если τ выбрано достаточно малым. Эта сложность оптимальна по зависимости от точности по [17].

Чтобы получить ту же сложность в случае, когда $\Delta > 0$, нам нужно выбрать подходящее значение τ и убедиться, что Δ достаточно мало. Таким образом, слагаемые $2M_2\tau$ и $\frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi$ в (3.1) должны иметь порядок ε. Эти условия также делают пренебрежимо малым член, зависящий от τ, в σ_q. Следовательно,

$$\text{когда} \quad \tau = \frac{\varepsilon}{M_2}\text{и}\Delta \leq \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}, \text{ имеем } T = \left(\frac{R_0\sqrt{d}a_qM_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}.$$

В противном случае, когда $\Delta > \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}$, скорость сходимости ухудшается. Как мы видим в (3.2), в этом случае мы не можем гарантировать точность меньше, чем $\sqrt{M_2\sqrt{d}\Delta\mathcal{D}_\Psi}$. Кроме того, количество итераций, необходимых, чтобы сделать другие слагаемыми меньшими, чем ε, составляет $T = O\left(\frac{\sqrt{M_2R_0da_q\Delta}}{\varepsilon}\right)^{\frac{2(1+\kappa)}{\kappa}}$. Это в два раза хуже, чем $O(\varepsilon^{-\frac{\kappa+1}{\kappa}})$, полученное при малом Δ .

Зависимость верхних оценок от q и d. В алгоритме 1 мы можем свободно выбирать $p \in [1{,}2]$ и Ψ_p, которые в зависимости от выпуклого компакта $\mathcal{X}$ приводят к различным значениям $\mathcal{D}_\Psi, R_0, a_q$. Желательно уменьшить $a_q, \mathcal{D}_\Psi$ одновременно, что позволило бы увеличить максимальный уровень шума Δ и быстрее сходиться без изменения скорости по (3.1). Однако в отличие от хорошо изученного SMD-алгоритма с сильно выпуклыми прокс-функциями Ψ_p, существует лишь несколько примеров эффективного выбора равномерно выпуклых прокс-функций Ψ_p, приведенных в [13].

Глава 4

Безградиентный Алгоритм с Клиппингом

4.0.1 Алгоритм и Теорема Сходимости

Альтернативный подход к работе с шумом с тяжелыми хвостами в стохастической оптимизации основан на методе клиппирования градиента, см., например, [18]. При заданной константе клиппирования $c > 0$ оператор клиппирования, применяемый к вектору g, определяется выражением

$$\hat{g} = \frac{g}{\|g\|} \min(\|g\|, c).$$

Клиппированный градиент имеет несколько полезных свойств для дальнейших доказательств, они приведены в лемме ниже.

Лемма 4.0.1. *Для $c > 0$ и случайного вектора $g = g(x,\xi,\mathbf{e})$,мы определяем $\hat{g} = \frac{g}{\|g\|_q} \min(\|g\|_q, c)$. Тогда мы имеем*

1.
$$\|\hat{g} - \mathbb{E}[\hat{g}]\|_q \leq 2c. \tag{4.1}$$

2. *Если дополнительно* $\mathbb{E}[\|g(x,\xi,\mathbf{e})\|_q^{1+\kappa}] \leq \sigma_q^{1+\kappa}$, *тогда*

(a)

$$\mathbb{E}[\|\hat{g}\|_q^2] \leq \sigma_q^{1+\kappa} c^{1-\kappa}, \tag{4.2}$$

(b)

$$\mathbb{E}[\|\hat{g} - \mathbb{E}[\hat{g}]\|_q^2] \leq 4\sigma_q^{1+\kappa} c^{1-\kappa}, \tag{4.3}$$

(c)

$$\|\mathbb{E}[g] - \mathbb{E}[\hat{g}]\|_q \leq \frac{\sigma_q^{1+\kappa}}{c^\kappa}. \tag{4.4}$$

Константа клиппирования c позволяет найти компромисс между более быстрой сходимостью из-за ограниченного второго момента $\hat{g}$ и смещения $\|\mathbb{E}[\hat{g} - g]\|$, когда $c \to 0$. Алгоритм, реализующий эту идею в наших условиях, представлен ниже.

Algorithm 2 Zeroth-Order Clipping Algorithm

procedure ZERO CLIP(Количество итераций T, размер шага ν, констакнта клиппирования c, прокс-функция Ψ_p, константа сглаживания τ)
  $x_0 \leftarrow \arg\min_{x \in \mathcal{X}} \Psi_p(x)$
  for $k = 0, 1, \ldots, T-1$ **do**
    Независимо отсемплировать $\mathbf{e}_k \sim \text{Uniform}(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\})$
    Независимо отсемплировать ξ_k
    Вычислить $g_{k+1} = \frac{d}{2\tau}(\phi(x_k + \tau\mathbf{e}_k, \xi_k) - \phi(x_k - \tau\mathbf{e}_k, \xi_k))\mathbf{e}_k$
    Клиппировать $\hat{g}_{k+1} = \frac{g_{k+1}}{\|g_{k+1}\|_q} \min(\|g_{k+1}\|_q, c)$
    Вычислить $y_{k+1} \leftarrow \nabla(\Psi_p^*)(\nabla\Psi_p(x_k) - \nu\hat{g}_{k+1})$
    Вычислить $x_{k+1} \leftarrow \arg\min_{x \in \mathcal{X}} D_{\Psi_p}(x, y_{k+1})$
  end for
  return $\overline{x}_T \leftarrow \frac{1}{T}\sum_{k=0}^{T-1} x_k$
end procedure

Следующий результат дает скорость сходимости для вышеуказанного алгоритма с точки зрения ожидания разрыва субоптимальности.

Теорема 4.0.2. *Пусть функция f, удовлетворяющая Предположениям 1, 2, 3, $q \in [2, \infty]$, количество итераций T, константа сглаживания $\tau > 0$*

заданы заранее. Выберем 1-сильно выпуклую функцию по p-норме прокс-функцию $\Psi_p(x)$. *Установим размер шага* $\nu = \left(\frac{R_0^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}}\right)^{\frac{1}{1+\kappa}}$ *с* σ_q *из Леммы 2.0.2, расстоянием от начальной точки* x_0 *до решения* x^* $R_0^{\frac{1+\kappa}{\kappa}} \stackrel{def}{=} \frac{1+\kappa}{\kappa} D_{\Psi_p}(x^*, x_0)$ *и диаметром компакта* $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \stackrel{def}{=} \frac{1+\kappa}{\kappa} \sup\limits_{x,y\in\mathcal{X}} D_{\Psi_p}(x,y)$. *После установим константу клиппирования* $c = \frac{2\kappa\mathcal{D}_\Psi}{(1-\kappa)\nu}$. *Пусть* $\overline{x}_T$ *результат работы Алгоритма 2 с заданными выше параметрами.*

1. *Тогда имеем следующую оценку*

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \tag{4.5}$$

   где $\sigma_q^{1+\kappa} = 2^\kappa\left(\frac{\sqrt{d}}{2^{1/4}}a_qM_2\right)^{1+\kappa} + 2^\kappa\left(\frac{da_q\Delta}{\tau}\right)^{1+\kappa}$.

2. *К тому же, с оптимальным* $\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}da_q\Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$ *имеем*

$$\begin{aligned}\mathbb{E}[f(\overline{x}_T)] - f(x^*) \;&\leq\; \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}da_q\Delta}{T^{\frac{\kappa}{(1+\kappa)}}}}\\ &+\; \frac{2\sqrt{d}a_qM_2R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}}{T^{\frac{\kappa}{1+\kappa}}}.\end{aligned} \tag{4.6}$$

Набросок доказательства Теоремы 4.0.2. Доказательство основано на Теореме 2.0.4 и неравенстве (2.9) для 1-сильно выпуклых Ψ_p, которые дают оценку

$$\underbrace{\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\langle\hat{g}_{k+1}, x_k - x^*\rangle\right]}_{①} \leq \mathbb{E}\left[\frac{1}{2}\frac{R_0^2}{\nu T}\right] + \underbrace{\mathbb{E}\left[\frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}\|\hat{g}_{k+1}\|_q^2\right]}_{②}. \tag{4.7}$$

① слагаемое в (4.7) из-за выпуклости и аппроксимационных свойств $\hat{f}_\tau(x)$ из Леммы 2.0.1, неравенства концентрации из Леммы 6.1.6 и свойств клиппированных векторов из Леммы 4.0.1 можно ограничить сверху

$$
(1) \geq \mathbb{E}[f(\overline{x}_T)] - f(x^*) - 2M_2\tau - \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi - \frac{\mathcal{D}_\Psi \sigma_q^{1+\kappa}}{c^\kappa}.
$$

(2) слагаемое в (4.7) можно ограничить по Лемме 4.0.1 как

$$
(2) \leq \frac{\nu}{2} c^{1-\kappa}\sigma_q^{1+\kappa}.
$$

Объединяя все оценки вместе, мы получим

$$
\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{1}{2}\frac{R_0^2}{\nu T} + \frac{\nu}{2}\sigma_q^{1+\kappa}c^{1-\kappa} + \left(\frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi.
$$

Далее мы выбираем константу клиппирования равную $c = \frac{2\kappa\mathcal{D}_\Psi}{(1-\kappa)\nu}$. Остаётся только выбрать размер шага $\nu = \left(\frac{R_0^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}}\right)^{\frac{1}{1+\kappa}}$ и константу сглаживания τ. $\square$

Для полного доказательства мы отсылаем читателя в Главу 6.3.

Следующий результат является более сильным и дает скорость сходимости для вышеуказанного алгоритма с точки зрения разрыва субоптимальности с высокой вероятностью. Однако это приводит к дополнительному коэффициенту $\log\frac{1}{\delta}$, где δ — желаемый уровень достоверности. Мы используем $\tilde{O}(\cdot)$-обозначение, чтобы скрыть полиномиальные множители $\log\frac{1}{\delta}$.

Теорема 4.0.3. *Пусть функция f, удовлетворяющая Предположениям 1, 2, 3, $q \in [2, \infty]$, количество итераций T, константа сглаживания $\tau > 0$ заданы заранее. Выберем 1-сильно выпуклую функцию по p-норме прокс-функцию $\Psi_p(x)$. Установим константу клиппирования равную $c = T^{\frac{1}{(1+\kappa)}}\sigma_q$ с σ_q из Леммы 2.0.2. Установим размер шага $\nu = \frac{\mathcal{D}_\Psi}{c}$ с диаметром компакта $\mathcal{D}_\Psi^2 \overset{def}{=} 2\sup\limits_{x,y\in\mathcal{X}} D_{\Psi_p}(x,y)$. Пусть $\overline{x}_T$ результат работы Алгоритма 2 с заданными выше параметрами.*

1. *Тогда с вероятностью не менее $1-\delta$ мы имеем оценку*

$$f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \frac{\Delta\sqrt{d}}{\tau}\mathcal{D}_\Psi + \tilde{O}\left(\frac{\mathcal{D}_\Psi \sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right), \tag{4.8}$$

где $\sigma_q^{1+\kappa} = 2^\kappa \left(\frac{\sqrt{d}}{2^{1/4}} a_q M_2\right)^{1+\kappa} + 2^\kappa \left(\frac{d a_q \Delta}{\tau}\right)^{1+\kappa}$.

2. *К тому же, с оптимальным* $\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4\mathcal{D}_\Psi d a_q \Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$ *имеем*

$$f(\overline{x}_T) - f(x^*) = \tilde{O}\left(\sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2\mathcal{D}_\Psi d a_q \Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{2\sqrt{d} a_q M_2 \mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}\right). \tag{4.9}$$

Набросок доказательства Теоремы 4.0.3. Для ограничения величин с вероятностью не менее $1-\delta$ воспользуемся классическим неравенством Бернштейна для суммы мартингальных разностей (т.е. $\mathbb{E}[X_i|X_{j<i}] = 0, \forall i \geq 1$) (Лемма 6.4.1) и для суммы квадратов случайных величин (Лемма 6.4.2).

Доказательство основано на Теореме 2.0.4 и неравенстве (2.9) для 1-сильно выпуклых Ψ_p, которые дают оценку

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1}, x_k - x^*\rangle \leq \frac{1}{2}\frac{R_0^2}{\nu T} + \underbrace{\frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}\|\hat{g}_{k+1}\|_q^2}_{1}. \tag{4.10}$$

Добавив $\pm\mathbb{E}_{|\leq k}[\hat{g}_{k+1}]$ и $\pm\hat{f}_\tau(x_k)$ к левой части (4.10), мы получим

$$
\begin{aligned}
\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1}, x_k - x^*\rangle = & \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1} - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle}_{2} \\
& + \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}\langle \mathbb{E}_{|\leq k}[\hat{g}_{k+1}] - \nabla\hat{f}_\tau(x_k), x_k - x^*\rangle,}_{3} \\
& + \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla\hat{f}_\tau(x_k), x_k - x^*\rangle\,.}_{4}
\end{aligned}
$$

Мы ограничим 1 слагаемое в (4.10), используя Лемму 6.4.2, а 2 ограничим как сумма мартингальных разностей через Лемму 6.4.1:

$$
1 = \tilde{O}\left(\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa} + \frac{1}{T}c^2\right)
$$

$$
2 = \tilde{O}\left(\frac{4c\mathcal{D}_\Psi}{T} + \frac{\sqrt{4\sigma_q^{1+\kappa}c^{1-\kappa}}}{\sqrt{T}}\mathcal{D}_\Psi^2\right).
$$

Далее мы ограничим 4 с помощью выпуклости $\hat{f}_\tau(x)$ из Леммы 2.0.1, 3, применив Лемму6.1.6и свойства климппированного вектора из Леммы 4.0.1:

$$
3 \leq \left(\frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi,
$$

$$
4 \geq f(\overline{x}_T) - f(x^*) - 2M_2\tau.
$$

Объединяя все оценки вместе, мы получим

$$f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \left(\frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi + \frac{1}{2}\frac{R_0^2}{\nu T}$$
$$+ \tilde{O}\left(\frac{\nu}{2}\sigma_q^{1+\kappa}c^{1-\kappa} + \frac{\nu}{2}\frac{1}{T}c^2 + \frac{4c\mathcal{D}_\Psi}{T} + \frac{\sqrt{4\sigma_q^{1+\kappa}c^{1-\kappa}}}{\sqrt{T}}\mathcal{D}_\Psi^2\right).$$

Далее мы выберем размер шага $\nu = \frac{\mathcal{D}_\Psi}{c}$, константу клиппирования $c = T^{\frac{1}{(1+\kappa)}}\sigma_q$, константу сглаживания τ и закончим доказательство. □

Для полного доказательства мы отсылаем читателя в Главу 6.4.

4.0.2 Обсуждение

В этом обсуждении мы сосредоточимся на оценках с высокой вероятностью, данных в теореме 4.0.3. То же самое справедливо и для результата теоремы 4.0.2 с точностью до исключения фактора $\log\frac{1}{\delta}$.

Максимальный уровень допустимого враждебного шума. Пусть $\varepsilon > 0$ — желаемая точность по значению функции, т. е. с вероятностью не менее $1-\delta$ имеем $f(\overline{x}_T) - f(x^*) \leq \varepsilon$.

В теореме 4.0.3, если нет враждебного шума, т.е. $\Delta = 0$, то количество итераций T для достижения этой точности определяется выражением $T = \tilde{O}\left(\left(\frac{\mathcal{D}_\Psi\sqrt{d}a_qM_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}\right)$, когда $\tau \to 0$. Эта скорость является оптимальной согласно [17].

Чтобы сохранить одинаковую скорость при $\Delta > 0$, слагаемые $2M_2\tau$ и $\frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi$ должно быть порядка ε. Эти условия также делают пренебрежимо малым член, зависящий от τ, в σ_q. Следовательно,

$$\text{когда}\quad \tau = \frac{\varepsilon}{M_2}\text{и}\Delta \leq \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi} \Rightarrow T = \tilde{O}\left(\left(\frac{\mathcal{D}_\Psi\sqrt{d}a_qM_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}\right).$$

В противном случае, когда $\Delta > \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}$, скорость сходимости ухудшается. Как и в случае с Robust SMD, мы не можем добиться точности меньше, чем $\sqrt{M_2\sqrt{d}\Delta\mathcal{D}_\Psi}$. Скорость сходимости к этой границе определяется выражением $T = \tilde{O}\left(\left(\frac{M_2\mathcal{D}_\Psi d a_q \Delta}{\varepsilon^2}\right)^{\frac{1+\kappa}{\kappa}}\right)$, что вдвое хуже, чем $\tilde{O}\left(\left(\frac{\mathcal{D}_\Psi\sqrt{d}a_q M_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}\right)$.

Рекомендации по выбору Ψ_p. В Алгоритме 2 мы можем свободно выбирать $p \in [1,2]$ и Ψ_p, которые в зависимости от выпуклого компакта $\mathcal{X}$ изменят $\mathcal{D}_\Psi, R_0, a_q$. Основная задача состоит в одновременном уменьшении $a_q, \mathcal{D}_\Psi$, что позволит нам увеличить максимальный шум Δ и быстрее сходиться без изменения скорости в соответствии с (4.8).

Далее мы обсудим некоторые стандартные множества $\mathcal{X}$ и прокс-функции Ψ_p, взятые из [16]. Два главных режима даны ниже

1. Ball setup:

$$p = 2, \Psi_p(x) = \frac{1}{2}\|x\|_2^2, \tag{4.11}$$

2. Entropy setup:

$$p = 1, \Psi_p(x) = (1+\gamma)\sum_{i=1}^{d}(x_i + \gamma/d)\log(x_i + \gamma/d), \gamma > 0. \tag{4.12}$$

Мы обозначаем единичный шар через $B^p = \{x \in \mathbb{R}^d : \|x\|_p \leq 1\}$ и стандартный симплекс через $\Delta_d^+ = \{x \in \mathbb{R}^d : x \geq 0, \sum_i x_i = 1\}$. По Лемме 2.0.2 константа $a_q = d^{\frac{1}{q}-\frac{1}{2}}\min\{\sqrt{32\ln d - 8}, \sqrt{2q-1}\}$. В следующих таблицах собрана сложность $T^{\frac{\kappa}{1+\kappa}}$ и максимально допустимый уровень шума Δ с точностью до $O\left(\log\frac{1}{\delta}\right)$ для каждого режима (строка) и множества (столбец).

Из этих таблиц видно, что для $\mathcal{X} = \Delta_d^+$ или B^1 Entropy setup предпочтительнее, а Ball setup допускает уровень шума Δ до $\sqrt{\ln d}$ раз больше. Между тем, для $\mathcal{X} = B^2$ или B^∞ Ball setup лучше с точки зрения скорости сходимости и устойчивости.

Таблица 4.1: $T^{\frac{\kappa}{1+\kappa}}$ с точностью $O\left(\log\frac{1}{\delta}\right)$ для Алгоритма 2

	Δ_d^+	B^1	B^2	B^∞
Ball	$\sqrt{d}M_2/\varepsilon$	$\sqrt{d}M_2/\varepsilon$	$\sqrt{d}M_2/\varepsilon$	dM_2/ε
Entropy	$\ln dM_2/\varepsilon$	$\ln dM_2/\varepsilon$	$\sqrt{d}\ln dM_2/\varepsilon$	$d\ln dM_2/\varepsilon$

Таблица 4.2: Максимально допустимый уровень шума Δ с точностью до $O(1)$ для Алгоритма 2

	Δ_d^+	B^1	B^2	B^∞
Ball	$\varepsilon^2/(\sqrt{d}M_2)$	$\varepsilon^2/(\sqrt{d}M_2)$	$\varepsilon^2/(\sqrt{d}M_2)$	$\varepsilon^2/(dM_2)$
Entropy	$\varepsilon^2/(\sqrt{d\ln d}M_2)$	$\varepsilon^2/(\sqrt{d\ln d}M_2)$	$\varepsilon^2/(d\sqrt{\ln d}M_2)$	$\varepsilon^2/(\sqrt{d^3\ln d}M_2)$

Оптимальность оценки в терминах d. В гладком случае для того, чтобы оценить градиент функции, достаточно использовать $d+1$ значений функции (см., например, [19]). Для стохастических методов первого порядка, оптимальное количество вызовов оракула первого порядка пропорционально $\varepsilon^{-\frac{1+\kappa}{\kappa}}$, поэтому для методов нулевого порядка можно ожидать $d\varepsilon^{-\frac{1+\kappa}{\kappa}}$. В этой работе мы получаем оценку $\left(\sqrt{d}/\varepsilon\right)^{\frac{1+\kappa}{\kappa}}$, которая совпадает только при $\kappa = 1$.

Оптимальна ли эта оценка?

Для гладких задач стохастической выпуклой оптимизации с $(d+1)$-точечным оракулом нулевого порядка ответ отрицательный, и оптимальная оценка количества значений функции равна $\sim d\varepsilon^{-\frac{1+\kappa}{\kappa}}$.

Сравнение Алгоритма с Клиппингом и Алгоритма с устойчивым SMD. Хотя обе теоремы о сходимости 3.0.1 и 4.0.2, 4.0.3 для алгоритмов 1, 2 соответственно дают одинаковые оценки, Алгоритм 2 гораздо более гибкий за счет выбора прокс-функций Ψ_p и возможности эффективной работы с разными множествами. Кроме того, алгоритм 2 гарантирует сходимость с высокой вероятностью. Однако на практике его сходимость резко зависит от константы клиппирования c, которую необходимо тщательно выбирать, а также от размера шага ν и константы сглаживания τ.

Глава 5

Безградиентный Алгоритм с Рестартами

5.0.1 Алгоритм и Теорема Сходимости

В этой главе предполагается, что целевая функция удовлетворяет условию острого минимума или r-growth из работы [20]. В этом случае алгоритмы оптимизации можно ускорить с помощью техники рестартов из работы [21].

Предположение 4. *Фукнуция f удовлетворяет условию r-growth, если существует $r \geq 1$ и $\mu_r \geq 0$ такие, что для любых x*

$$\frac{\mu_r}{2}\|x - x^*\|_p^r \leq f(x) - f(x^*),$$

где x^* – решение проблемы.

В частности, условие μ-сильной выпуклости по ℓ_p-норме является условие 2-growth. Заметим, что техника рестартов работает, если Δ достаточно мала, чтобы сохранить оптимальность алгоритмов 1 и 2. Общая схема алгоритма с рестартами представлена ниже.

Algorithm 3 Zeroth-Order Restart Algorithm

procedure Zeroth-Order Restart(Алгоритм $\mathcal{A}$, количество рестартов N, последовательность количества итераций $\{T_k\}_{k=1}^N$, последовательность констант сглаживания $\{\tau_k\}_{k=1}^N$, последовательность размеров шагов $\{\nu_k\}_{k=1}^N$, последовательность констант клиппирования $\{c_k\}_{k=1}^N$(при необходимости), прокс-функция Ψ_p)
  Установить $x_0 \leftarrow \arg\min\limits_{x\in\mathcal{X}} \Psi_p(x)$ или выбрать случайно
  for $k = 0, 1, \dots, N$ **do**
    Установить параметры $\nu_k, (c_k), \Psi_p, \tau_k$ Алгоритма $\mathcal{A}$
    Вычислить T_k итераций Алгоритма $\mathcal{A}$ с начальной точкой x_0 и получить на выходе x_{final}
    $x_0 \leftarrow x_{\text{final}}$
  end for
  return x_{final}
end procedure

Соответствующие теоремы, обобщающие полученные выше результаты для алгоритмов 1, 2, состоят в следующем. Мы используем обозначение $\tilde{O}(\cdot)$ ниже, чтобы скрыть полиномиальные множители $\log d$.

Теорема 5.0.1. *Пусть функция f удовлетворяет Предположениям 1, 2. Также $\varepsilon > 0$ является зафиксированной точностью и r-growth Предположение 4 верно с $r \geq \frac{1+\kappa}{\kappa}$.*

Посчитаем $R_0 \overset{def}{=} \sup_{x,y\in\mathcal{X}} \left(\frac{1+\kappa}{\kappa} D_{\Psi_p}(x,y)\right)^{\frac{\kappa}{1+\kappa}}$ и $R_k = R^0/2^k$.

Установим количество рестартов $N = \tilde{O}\left(\frac{1}{r}\log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)\right)$, последовательность количества итераций $\{T_k\}_{k=1}^N = \left\{\tilde{O}\left(\left[\frac{\sigma_q 2^{(1+r)}}{\mu_r R_k^{r-1}}\right]^{\frac{1+\kappa}{\kappa}}\right)\right\}_{k=1}^N$, последовательность констант сглаживания $\{\tau_k\}_{k=1}^N = \left\{\frac{\sigma_q R_k}{M_2 T_k^{\frac{\kappa}{1+\kappa}}}\right\}_{k=1}^N$ и последовательность размера шагов $\{\nu_k\}_{k=1}^N = \left\{\frac{R_k^{1/\kappa}}{\sigma_q} T_k^{-\frac{1}{1+\kappa}}\right\}_{k=1}^N$, где σ_q получено в Лемме 2.0.2. Наконец, пусть выполняется Предположение 3 с

$$\Delta_k = \tilde{O}\left(\frac{\mu_r^2 R_0^{(2r-1)}}{M_2\sqrt{d}} \frac{1}{2^{k(2r-1)}}\right), \quad 1 \leq k \leq N.$$

Если x_{final} является результатом работы Алгоритма 3 с базовым Алго-

ритмом с Устойчивым SMD 1 как $\mathcal{A}$ и с указанными выше параметрами, то

$$\mathbb{E}[f(x_{\text{final}})] - f(x^*) \le \varepsilon,$$

а общее количество итераций

$$T = \tilde{O}\left(\left[\frac{a_q M_2 \sqrt{d}}{\mu_r^{1/r}} \cdot \frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right), \quad a_q \overset{def}{=} d^{\frac{1}{q}-\frac{1}{2}} \min\{\sqrt{32\ln d - 8}, \sqrt{2q-1}\}.$$

При этом на последнем рестарте Δ должно быть

$$\Delta_N = \tilde{O}\left(\frac{\mu_r^{1/r}}{M_2\sqrt{d}} \varepsilon^{(2-1/r)}\right).$$

Теорема 5.0.2. [1] *Пусть функция f удовлетворяет Предположениям 1, 2. Также $\varepsilon > 0$ является зафиксированной точностью и r-growth Предположение 4 верно с $r \ge 2$ для оценки по мат. ожиданию или $r \ge 1$ для оценки с высокой вероятностью. Посчитаем $R_0 \overset{def}{=} \sup_{x,y\in\mathcal{X}} \left(2D_{\Psi_p}(x,y)\right)^{\frac{1}{2}}$ и $R_k = R^0/2^k$. Установим количество рестартов $N = \tilde{O}\left(\frac{1}{r}\log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)\right)$, последовательность количества итераций $\{T_k\}_{k=1}^N = \left\{\tilde{O}\left(\left[\frac{\sigma_q 2^{(1+r)}}{\mu_r R_k^{r-1}}\right]^{\frac{1+\kappa}{\kappa}}\right)\right\}_{k=1}^N$, последовательность констант сглаживания $\{\tau_k\}_{k=1}^N = \left\{\frac{\sigma_q R_k}{M_2 T_k^{\frac{\kappa}{1+\kappa}}}\right\}_{k=1}^N$, последовательность констант клиппирования $\{c_k\}_{k=1}^N = \left\{T_k^{\frac{1}{(1+\kappa)}}\sigma_q\right\}_{k=1}^N$ и последовательность размера шагов $\{\nu_k\}_{k=1}^N = \left\{\frac{R_k}{c_k}\right\}_{k=1}^N$, где σ_q получено в Лемме 2.0.2. Наконец, пусть выполняется Предположение 3 с*

$$\Delta_k = \tilde{O}\left(\frac{\mu_r^2 R_0^{(2r-1)}}{M_2\sqrt{d}} \frac{1}{2^{k(2r-1)}}\right), \quad 1 \le k \le N.$$

[1] В этой теореме $\tilde{O}(\cdot)$ обозначает множитель $\log d$ для оценки по мат.ожиданию и множители $\log d, \log\frac{1}{\delta}$ для оценки с высокой вероятностью. Более явные формулы приведены в полном доказательстве.

Если x_{final} является результатом работы Алгоритма 3 с базовым Алгоритмом с Клиппингом 2 как $\mathcal{A}$ и с указанными выше параметрами, то

$$\mathbb{E}[f(x_{\text{final}})] - f(x^*) \leq \varepsilon,$$

или с вероятностью не менее $1-\delta$

$$f(x_{\text{final}}) - f(x^*) \leq \varepsilon.$$

Общее число итераций равно

$$T = \tilde{O}\left(\left[\frac{a_q M_2\sqrt{d}}{\mu_r^{1/r}} \cdot \frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right), \quad a_q \overset{def}{=} d^{\frac{1}{q}-\frac{1}{2}} \min\{\sqrt{32\ln d - 8}, \sqrt{2q-1}\},$$

а на последнем рестарте Δ равен

$$\Delta_N = \tilde{O}\left(\frac{\mu_r^{1/r}}{M_2\sqrt{d}}\varepsilon^{(2-1/r)}\right).$$

5.0.2 Обсуждение

Максимальный уровень допустимого враждебного шума. В отличие от Алгоритмов с Клиппингом или Устойчивым SMD, Алгоритм с Рестартами и Предположение r-growth гарантируют более высокий максимальный порог для Δ, а именно,

$$\text{Алгоритм 1 или 2}: \qquad \Delta = \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi},$$
$$\text{Алгоритм3} \;: \qquad \Delta = \tilde{O}\left(\frac{\mu_r^{1/r}}{M_2\sqrt{d}}\varepsilon^{(2-1/r)}\right).$$

Более того, этот порог не зависит от множества $\mathcal{X}$, а фактор $\frac{1}{\sqrt{d}}$ является наилучшим. Кроме того, вначале Δ_k может быть намного больше и начинает уменьшаться как $\Delta_k = \frac{\Delta_1}{2^{k(2r-1)}}$ только при последующих перезапусках для достижения необходимой точность.

Отметим также, что чем меньше r, тем выше порог.

Зависимость от q,d, ε Опять же, в отличие от Алгоритмов с Клиппингом или Устойчивым SMD, Алгоритм с Рестартами и Предположение r-growth гарантируют лучшую оракульную сложность в зависимости от ε. Ниже мы приводим оценки ожиданий

$$\text{Алгоритм 1 или 2 :} \qquad T = \left[\frac{\mathcal{D}_\Psi\sqrt{d}a_q M_2}{\varepsilon}\right]^{\frac{1+\kappa}{\kappa}},$$

$$\text{Алгоритм 3 :} \qquad T = \tilde{O}\left(\left[\frac{a_q M_2\sqrt{d}}{\mu_r^{1/r}}\cdot\frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right).$$

Кроме того, в Алгоритме с Рестартами общее количество итераций зависит только от a_q, а максимальный порог Δ вообще не зависит от q,$\mathcal{X}$. Таким образом, имеет смысл использовать Entropy setup, определенную в (4.12), с базовым Алгоритмом с Клиппингом, чтобы снизить a_q и оставить только фактор $\log d$ в оценке T.

Глава 6

Заключение

Для d-мерной безградиентоной оптимизации с двухточечным оракулом и тяжелыми хвостами мы предлагаем алгоритмы на основе клиппирования и на основе устойчивого стохастического зеркального спуска. Алгоритм на основе клиппирования имеет оракульную сложность пропорциональную $\left(\sqrt{d}/\varepsilon\right)^{\frac{1+\kappa}{\kappa}}$ в терминах высокой вероятности и мат. ожидания. Он также имеет максимально допустимый уровень враждебного шума $\lesssim \varepsilon^2/\sqrt{d}$.

Кроме того, мы обобщаем эти результаты на задачи с функцией, удовлетворяющей условию r-growth, к которым относятся сильно выпуклые задачи и задачи с острым минимумом. Мы используем технику рестартов для алгоритма на основе клиппирования, описанного выше. В этом случае мы получаем оракульную сложность $\sim \left(\sqrt{d}/\varepsilon^{\frac{(r-1)}{r}}\right)^{\frac{1+\kappa}{\kappa}}$ в терминах высокой вероятности и мат. ожидания, а также максимально допустимый уровень враждебного шума $\lesssim \varepsilon^{(2-1/r)}/\sqrt{d}$.

Мы считаем, что скорость сходимости может быть улучшена с помощью следующих возможных модификаций:

1. Использовать другую стратегию семплирования для оценки g_k, а именно равномерное семплирование с единичной ℓ_1-сферы $\{\mathbf{e} : \|\mathbf{e}\|_1 = 1\}$, см., например, [22], [23].

2. Использовать другие предположения на враждебный шум, а именно предположение о липшицевой непрерывности

$$|\delta(\mathrm{x}_1) - \delta(\mathrm{x}_2)| \leq M\|x_1 - x_2\|_2, \qquad \forall x_1, x_2 \in \mathcal{X}$$

см., например, [9].

3. Использовать адаптивные стратегии и эвристические методы для выбора входных параметров алгоритма, таких как размер шага ν, константа сглаживания τ и т.д. На практике эти константы трудно оценить.

Литература

[1] *Spall, James C.* Introduction to stochastic search and optimization: estimation, simulation, and control / James C Spall. — Chichester: John Wiley & Sons, 2005.

[2] *Conn, Andrew R.* Introduction to derivative-free optimization / Andrew R Conn, Katya Scheinberg, Luis N Vicente. — Montreal: SIAM, 2009.

[3] Optimal rates for zero-order convex optimization: The power of two function evaluations / John C Duchi, Michael I Jordan, Martin J Wainwright, Andre Wibisono // *IEEE Transactions on Information Theory.* — 2015. — Vol. 61, no. 5. — Pp. 2788–2806.

[4] Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex / Alexander V Gasnikov, Anastasia A Lagunovskaya, Ilnura N Usmanova, Fedor A Fedorenko // *Automation and Remote Control.* — 2016. — Vol. 77. — Pp. 2018–2034.

[5] *Nesterov, Yurii.* Random gradient-free minimization of convex functions / Yurii Nesterov, Vladimir Spokoiny // *Foundations of Computational Mathematics.* — 2017. — Vol. 17. — Pp. 527–566.

[6] Stochastic online optimization. Single-point and multi-point non-linear multi-armed bandits. Convex and strongly-convex case / Alexander V Gasnikov, Ekaterina A Krymova, Anastasia A Lagunovskaya

et al. // *Automation and remote control.* — 2017. — Vol. 78. — Pp. 224–234.

[7] *Shamir, Ohad.* An optimal algorithm for bandit and zero-order convex optimization with two-point feedback / Ohad Shamir // *The Journal of Machine Learning Research.* — 2017. — Vol. 18, no. 1. — Pp. 1703–1713.

[8] *Bayandina, Anastasia Sergeevna.* Gradient-free two-point methods for solving stochastic nonsmooth convex optimization problems with small non-random noises / Anastasia Sergeevna Bayandina, Alexander V Gasnikov, Anastasia A Lagunovskaya // *Automation and Remote Control.* — 2018. — Vol. 79. — Pp. 1399–1408.

[9] Gradient-Free Optimization for Non-Smooth Minimax Problems with Maximum Value of Adversarial Noise / Darina Dvinskikh, Vladislav Tominin, Yaroslav Tominin, Alexander Gasnikov // *arXiv preprint arXiv:2202.06114.* — 2022.

[10] Algorithms of robust stochastic optimization based on mirror descent method / Alexander V Nazin, Arkadi S Nemirovsky, Alexandre B Tsybakov, Anatoli B Juditsky // *Automation and Remote Control.* — 2019. — Vol. 80. — Pp. 1607–1627.

[11] From low probability to high confidence in stochastic convex optimization / Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, Junyu Zhang // *The Journal of Machine Learning Research.* — 2021. — Vol. 22, no. 1. — Pp. 2237–2274.

[12] Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise / Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev et al. // *arXiv preprint arXiv:2106.05958.* — 2021.

[13] Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance / Nuri Mert Vural, Lu Yu, Krishna Balasubramanian et al. // Conference on Learning Theory / PMLR. — 2022. — Pp. 65–102.

[14] The power of first-order smooth optimization for black-box non-smooth problems / Alexander Gasnikov, Anton Novitskii, Vasilii Novitskii et al. // *arXiv preprint arXiv:2201.12289.* — 2022.

[15] *Nemirovskij, Arkadij Semenovič.* Problem complexity and method efficiency in optimization / Arkadij Semenovič Nemirovskij, David Borisovich Yudin. — 1983.

[16] *Ben-Tal, Aharon.* Lectures on modern convex optimization: analysis, algorithms, and engineering applications / Aharon Ben-Tal, Arkadi Nemirovski. — Philadelphia: SIAM, 2001.

[17] *Nemirovsky, AS.* Problem complexity and optimization method efficiency / AS Nemirovsky, DB Yudin // *M.: Nauka.* — 1979.

[18] Why are adaptive methods good for attention models? / Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit et al. // *Advances in Neural Information Processing Systems.* — 2020. — Vol. 33. — Pp. 15383–15393.

[19] Randomized gradient-free methods in convex optimization / Alexander Gasnikov, Darina Dvinskikh, Pavel Dvurechensky et al. // *arXiv preprint arXiv:2211.13566.* — 2022.

[20] *Shapiro, Alexander.* Lectures on stochastic programming: modeling and theory / Alexander Shapiro, Darinka Dentcheva, Andrzej Ruszczynski. — Philadelphia: SIAM, 2021.

[21] *Juditsky, Anatoli.* Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization / Anatoli Juditsky, Yuri Nesterov // *Stochastic Systems.* — 2014. — Vol. 4, no. 1. — Pp. 44–80.

[22] A gradient estimator via L1-randomization for online zero-order optimization with two point feedback / Arya Akhavan, Evgenii Chzhen, Massimiliano Pontil, Alexandre B Tsybakov // *arXiv preprint arXiv:2205.13910.* — 2022.

[23] Gradient-Free Federated Learning Methods with l_1 and l_2-Randomization for Non-Smooth Convex Stochastic Optimization Problems / Aleksandr Lobanov, Belal Alashqar, Darina Dvinskikh, Alexander Gasnikov // *arXiv preprint arXiv:2211.10783.* — 2022.

[24] *Beznosikov, Aleksandr.* Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem / Aleksandr Beznosikov, Abdurakhmon Sadiev, Alexander Gasnikov // Mathematical Optimization Theory and Operations Research: 19th International Conference, MOTOR 2020, Novosibirsk, Russia, July 6–10, 2020, Revised Selected Papers 19 / Springer. — 2020. — Pp. 105–119.

[25] *Gorbunov, Eduard A.* Stochastic Optimization with Heavy-Tailed Noise via Accelerated Gradient Clipping / Eduard A. Gorbunov, Marina Danilova, Alexander V. Gasnikov // *ArXiv.* — 2020. — Vol. abs/2005.10785.

[26] High-Probability Bounds for Stochastic Optimization and Variational Inequalities: the Case of Unbounded Variance / Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov et al. // *arXiv preprint arXiv:2302.00999.* — 2023.

[27] High Probability Convergence of Clipped-SGD Under Heavy-tailed Noise / Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, Huy Le Nguyen // *arXiv preprint arXiv:2302.05437.* — 2023.

[28] *Liu, Zijian.* Stochastic Nonsmooth Convex Optimization with Heavy-Tailed Noises / Zijian Liu, Zhengyuan Zhou // *arXiv preprint arXiv:2303.12277.* — 2023.

[29] *Gorbunov, EA.* On the Upper Bound for the Expectation of the Norm of a Vector Uniformly Distributed on the Sphere and the Phenomenon of Concentration of Uniform Measure on the Sphere. / EA Gorbunov, Evgeniya Alekseevna Vorontsova, Alexander Vladimirovich Gasnikov // *Mathematical Notes.* — 2019. — Vol. 106.

[30] *Ledoux, Michel.* The Concentration of Measure Phenomenon. Ed. by Peter Landweber et al. Vol. 89 / Michel Ledoux // *Mathematical Surveys and Monographs. Providence, Rhode Island: American Mathematical Society.* — 2005. — P. 181.

[31] Optimal mean estimation without a variance / Yeshwanth Cherapanamjeri, Nilesh Tripuraneni, Peter Bartlett, Michael Jordan // Conference on Learning Theory / PMLR. — 2022. — Pp. 356–357.

[32] *Nguyen, Ta Duy.* Improved Convergence in High Probability of Clipped Gradient Methods with Heavy Tails / Ta Duy Nguyen, Alina Ene, Huy L Nguyen // *arXiv preprint arXiv:2304.01119.* — 2023.

[33] *Zhang, Jiujia.* Parameter-free Regret in High Probability with Heavy Tails / Jiujia Zhang, Ashok Cutkosky // *arXiv preprint arXiv:2210.14355.* — 2022.

[34] *Gasnikov, Alexander Vladimirovich.* Universal method for stochastic composite optimization problems / Alexander Vladimirovich Gasnikov,

Yu E Nesterov // *Computational Mathematics and Mathematical Physics.* — 2018. — Vol. 58. — Pp. 48–64.

[35] *Lan, Guanghui.* Validation analysis of mirror descent stochastic approximation method / Guanghui Lan, Arkadi Nemirovski, Alexander Shapiro // *Mathematical programming.* — 2012. — Vol. 134, no. 2. — Pp. 425–458.

Приложение

6.1 Доказательства лемм

6.1.1 Общие результаты

Лемма 6.1.1. *1. Для любых $x,y \in \mathbb{R}^{d'}$ и $\kappa \in (0,1]$:*

$$\|x-y\|_q^{1+\kappa} \leq 2^\kappa \|x\|_q^{1+\kappa} + 2^\kappa \|y\|_q^{1+\kappa}, \tag{6.1}$$

2.

$$\forall x,y \geq 0, \kappa \in [0,1] : (x+y)^\kappa \leq x^\kappa + y^\kappa. \tag{6.2}$$

Доказательство. • По неравенству Йенсена для выпуклой функции $\|\cdot\|_q^{1+\kappa}$ с $1+\kappa > 1$ мы получаем

$$\|x-y\|_q^{1+\kappa} = 2^{1+\kappa}\|x/2 - y/2\|_q^{1+\kappa} \leq 2^\kappa \|x\|_q^{1+\kappa} + 2^\kappa \|y\|_q^{1+\kappa}.$$

• Утверждение 9 из работы [13].

□

Лемма 6.1.2. *Из Предположения 2 следует, что $f(x)$ является M_2 Липшецевой на $\mathcal{X}$.*

Доказательство. Для всех $x,y \in \mathcal{X}$

$$\begin{aligned}|f(x)-f(y)| &= |\mathbb{E}_\xi[f(x,\xi)-f(y,\xi)]| \overset{\text{нер. Йенсена}}{\leq} \mathbb{E}_\xi[|f(x,\xi)-f(y,\xi)|] \\ &\leq \mathbb{E}_\xi[M_2]\|x-y\|_2 \overset{\text{нер. Йенсена}}{\leq} \mathbb{E}_\xi[M_2^{(1+\kappa)}]^{\frac{1}{1+\kappa}}\|x-y\|_2 \\ &\leq M_2\|x-y\|_2. \end{aligned} \tag{6.3}$$

□

6.1.2 Сглаживание

Лемма 6.1.3. *Пусть $f(x)$ является M_2 Липшецевой функцией по норме $\|\cdot\|_2$. Если случайный вектор $\mathbf{e}$ равномерно равпределён на Евклидовой сфере и $\kappa \in (0,1]$, тогда*

$$\mathbb{E}_{\mathbf{e}}\left[(f(\mathbf{e})-\mathbb{E}_{\mathbf{e}}[f(\mathbf{e})])^{2(1+\kappa)}\right] \leq \left(\frac{bM_2^2}{d}\right)^{1+\kappa}, \quad b = \frac{1}{\sqrt{2}}.$$

Доказательство. Стандартный результат концентрации меры на евклидовой единичной сфере подразумевает, что $\forall t > 0$

$$Pr\left(|f(\mathbf{e})-\mathbb{E}_{\mathbf{e}}[f(\mathbf{e})]| > t\right) \leq 2\exp(-b'dt^2/M_2^2), \quad b' = 2 \tag{6.4}$$

(см. доказательство Предложения 2.10 и Следствия 2.6 из [30]). Из этого неравенства следует цепочка неравенств ниже

$$\mathbb{E}_{\mathbf{e}}\left[(f(\mathbf{e})-\mathbb{E}_{\mathbf{e}}[f(\mathbf{e})])^{2(1+\kappa)}\right] = \int\limits_{t=0}^{\infty} Pr\left(|f(\mathbf{e})-\mathbb{E}_{\mathbf{e}}[f(\mathbf{e})]|^{2(1+\kappa)} > t\right) dt$$
$$= \int\limits_{t=0}^{\infty} Pr\left(|f(\mathbf{e})-\mathbb{E}_{\mathbf{e}}[f(\mathbf{e})]| > t^{\frac{1}{2(1+\kappa)}}\right) dt$$
$$\leq \int\limits_{t=0}^{\infty} 2\exp\left(-b'dt^{\frac{1}{(1+\kappa)}}/M_2^2\right) dt \leq \left(\frac{bM_2^2}{d}\right)^{1+\kappa}.$$

□

Следующая лемма дает некоторые полезные сведения о концентрации меры на евклидовой единичной сфере.

Лемма 6.1.4. *Для* $q \geq 2, \kappa \in (0,1]$

$$\mathbb{E}_{\mathbf{e}}\left[\|\mathbf{e}\|_q^{2(1+\kappa)}\right] \leq a_q^{2(1+\kappa)} \stackrel{def}{=} d^{\frac{1}{q}-\frac{1}{2}} \min\{\sqrt{32\ln d - 8}, \sqrt{2q-1}\}.$$

Эта лемма является обобщением леммы из[29] для $\kappa < 1$.

Доказательство. Воспользуемся Леммой 1 из Теоремы 1 из [29].

1. Пусть e_k является k-ой компонентой вектора $\mathbf{e}$

$$\mathbb{E}\left[|e_k|^q\right] \leq \left(\frac{q-1}{d}\right)^{\frac{q}{2}}, \quad q \geq 2. \tag{6.5}$$

2. Для любых $x \in \mathbb{R}^d$ и $q_1 \geq q_2$

$$\|x\|_{q_1} \leq \|x\|_{q_2}, \tag{6.6}$$

Тогда

$$\mathbb{E}_{\mathbf{e}}\left[\|\mathbf{e}\|_q^{2(1+\kappa)}\right] = \mathbb{E}_{\mathbf{e}}\left[\left(\left(\sum_{k=1}^d |e_k|^q\right)^2\right)^{\frac{1+\kappa}{q}}\right].$$

С помощью неравенства Йенсена и одинаково распределённых e_k мы получаем

$$\mathbb{E}_{\mathbf{e}}\left[\left(\left(\sum_{k=1}^d |e_k|^q\right)^2\right)^{\frac{1+\kappa}{q}}\right] \leq \left(\mathbb{E}_{\mathbf{e}}\left[\left(\sum_{k=1}^d |e_k|^q\right)^2\right]\right)^{\frac{1+\kappa}{q}}.$$

Мы используем тот факт, что $\forall x_k \geq 0, k = \overline{1,d}$

$$d\sum_{k=1}^d x_k^2 \geq \left(\sum_{k=1}^d x_k\right)^2.$$

Из него следует, что

$$\left(\mathbb{E}_{\mathbf{e}}\left[\left(\sum_{k=1}^d |e_k|^q\right)^2\right]\right)^{\frac{1+\kappa}{q}} \leq \left(d\mathbb{E}_{\mathbf{e}}\left[\sum_{k=1}^d |e_k|^{2q}\right]\right)^{\frac{1+\kappa}{q}} = (d^2\mathbb{E}_{\mathbf{e}}[|e_k|^{2q}])^{\frac{1+\kappa}{q}}.$$

Используя (6.5) с $2q$, мы продолжаем цепочку предыдущих неравенств

$$(d^2\mathbb{E}_{\mathbf{e}}[|e_2|^{2q}])^{\frac{1+\kappa}{q}} \leq d^{\frac{2(1+\kappa)}{q}}\left(\frac{2q-1}{d}\right)^{1+\kappa} = \left(d^{\frac{2}{q}-1}(2q-1)\right)^{1+\kappa}.$$

Таким образом, по определению a_q и полученным оценкам заключаем

$$a_q = \sqrt{d^{\frac{2}{q}-1}(2q-1)}.$$

При фиксированном d и большом q можно получить более точную верхнюю оценку. Определим функцию $h_d(q)$ и найдем ее минимум при фиксированном d.

$$h_d(q) = \ln\left(\sqrt{d^{\frac{2}{q}-1}(2q-1)}\right) = \left(\frac{1}{q} - \frac{1}{2}\right)\ln(d) + \frac{1}{2}\ln(2q-1),$$

$$\frac{dh_d(q)}{dq} = \frac{-\ln(d)}{q^2} + \frac{1}{2q-1} = 0,$$
$$q^2 - 2\ln(d)q + \ln(d) = 0.$$

Когда $d \geq 3$ точка минимума q_0 лежит в $[2, +\infty)$

$$q_0 = (\ln d)\left(1 + \sqrt{1 - \frac{1}{\ln d}}\right), \qquad \ln d \leq q_0 \leq 2\ln d.$$

Когда $q \geq q_0$ мы получаем из (6.6)

$$\begin{aligned} a_q < a_{q_0} &= \sqrt{d^{\frac{2}{q_0}-1}(2q_0-1)} \leq d^{\frac{1}{\ln d}-\frac{1}{2}}\sqrt{4\ln d - 1} \\ &= \frac{e}{\sqrt{d}}\sqrt{4\ln d - 1} \leq d^{\frac{1}{q}-\frac{1}{2}}\sqrt{32\ln d - 8}, \end{aligned}$$

Следовательно, финальная оценка имеет вид

$$a_q = d^{\frac{1}{q}-\frac{1}{2}}\min\{\sqrt{32\ln d - 8}, \sqrt{2q-1}\}.$$

$\square$

Лемма 6.1.5. *Для случайного вектора* $\mathbf{e}$ *равномерно распределённого на евклидовой сфере* $\{\mathbf{e} \in \mathbb{R}^d : \|\mathbf{e}\|_2 = 1\}$ *и для любого* $r \in \mathbb{R}^d$, *мы имеем*

$$\mathbb{E}_{\mathbf{e}}[|\langle \mathbf{e}, r\rangle|] \leq \frac{\|r\|_2}{\sqrt{d}}.$$

Лемма 6.1.6. *Пусть* $g(x, \xi, \mathbf{e})$ *определён в* (2.4) *и* $\hat{f}_\tau(x)$ *определена в* (2.1).

Тогда при Предположении 3 верно следующее неравенство

$$\mathbb{E}_{\xi,\mathbf{e}}[\langle g(x,\xi,\mathbf{e}), r\rangle] \geq \langle \nabla \hat{f}_\tau(x), r\rangle - \frac{d\Delta}{\tau}\mathbb{E}_{\mathbf{e}}[|\langle \mathbf{e}, r\rangle|]$$

для любых $r \in \mathbb{R}^d$.

Доказательство. Напомним определение (2.4) оценки градиента g

$$g(x,\xi,\mathbf{e}) = \frac{d}{2\tau}(f(x+\tau\mathbf{e},\xi) + \delta(x+\tau\mathbf{e}) - f(x-\tau\mathbf{e},\xi) - \delta(x-\tau\mathbf{e}))\mathbf{e}.$$

Тогда, умножив g на произвольный вектор r и взяв полное мат. ожидание с обеих сторон, мы получим

$$\begin{aligned} \mathbb{E}_{\xi,\mathbf{e}}[\langle g(x,\xi,\mathbf{e}), r\rangle] &= \frac{d}{2\tau}\mathbb{E}_{\xi,\mathbf{e}}[\langle (f(x+\tau\mathbf{e},\xi) - f(x-\tau\mathbf{e},\xi))\mathbf{e}, r\rangle] \\ &+ \frac{d}{2\tau}\mathbb{E}_{\xi,\mathbf{e}}[\langle (\delta(x+\tau\mathbf{e}) - \delta(x-\tau\mathbf{e}))\mathbf{e}, r\rangle]. \end{aligned}$$

Для первого слагаемого мы используем тот факт, что $\mathbf{e}$ распределен симметрично

$$\begin{aligned} &\frac{d}{2\tau}\mathbb{E}_{\xi,\mathbf{e}}[\langle (f(x+\tau\mathbf{e},\xi) - f(x-\tau\mathbf{e},\xi))\mathbf{e}, r\rangle] = \frac{d}{\tau}\mathbb{E}_{\xi,\mathbf{e}}[\langle f(x+\tau\mathbf{e},\xi)\mathbf{e}, r\rangle] \\ &= \frac{d}{\tau}\mathbb{E}_{\mathbf{e}}[\langle \mathbb{E}_{\xi}[f(x+\tau\mathbf{e},\xi)]\mathbf{e}, r\rangle] = \frac{d}{\tau}\langle \mathbb{E}_{\mathbf{e}}[f(x+\tau\mathbf{e})\mathbf{e}], r\rangle. \end{aligned} \tag{6.7}$$

Используя Лемму 2.0.1 в (6.7), мы берём полное мат.ожидание

$$\frac{d}{\tau}\langle \mathbb{E}_{\mathbf{e}}[f(x+\tau\mathbf{e})\mathbf{e}], r\rangle = \langle \nabla \hat{f}_\tau(x), r\rangle.$$

Во торой слагаемом мы применяем Предположение 3

$$\frac{d}{2\tau}\mathbb{E}_{\xi,\mathbf{e}}[\langle (\delta(x+\tau\mathbf{e}) - \delta(x-\tau\mathbf{e}))\mathbf{e}, r\rangle] \geq -\frac{d\Delta}{\tau}\mathbb{E}_{\mathbf{e}}[|\langle \mathbf{e}, r\rangle|].$$

Складывая два слагаемых вместе, мы получаем нужный результат.

$\square$

Лемма 6.1.7. *При Предположениях 1, 2 и 3, для* $q \in [1, +\infty)$, *мы имеем*

$$\mathbb{E}_{\xi,\mathbf{e}}[\|g(x,\xi,\mathbf{e})\|_q^{1+\kappa}] \leq 2^\kappa \left(\frac{\sqrt{d}}{2^{1/4}} a_q M_2\right)^{1+\kappa} + 2^\kappa \left(\frac{d a_q \Delta}{\tau}\right)^{1+\kappa} = \sigma_q^{1+\kappa},$$

где $a_q \overset{def}{=} d^{\frac{1}{q}-\frac{1}{2}} \min\{\sqrt{32 \ln d - 8}, \sqrt{2q-1}\}$.

Доказательство. По определению (2.4) оценки градиента g получаем следующую цепочку неравенств

$$\begin{aligned}
\mathbb{E}_{\xi,\mathbf{e}}[\|g(x,\xi,\mathbf{e})\|_q^{1+\kappa}] &= \mathbb{E}_{\xi,\mathbf{e}} \left[\left\|\frac{d}{2\tau}(\phi(x+\tau\mathbf{e},\xi) - \phi(x-\tau\mathbf{e},\xi))\mathbf{e}\right\|_q^{1+\kappa}\right] \\
\overset{(6.1)}{\leq} \quad & 2^\kappa \left(\frac{d}{2\tau}\right)^{1+\kappa} \mathbb{E}_{\xi,\mathbf{e}} \left[\|\mathbf{e}\|_q^{1+\kappa} |f(x+\tau\mathbf{e},\xi) - f(x-\tau\mathbf{e},\xi)|^{1+\kappa}\right] \qquad (6.8) \\
+ \quad & 2^\kappa \left(\frac{d}{2\tau}\right)^{1+\kappa} \mathbb{E}_{\xi,\mathbf{e}} \left[\|\mathbf{e}\|_q^{1+\kappa} |\delta(x+\tau\mathbf{e}) - \delta(x-\tau\mathbf{e})|^{1+\kappa}\right]. \qquad (6.9)
\end{aligned}$$

Сначала разберемся со слагаемым (6.8). Добавляя $\pm\alpha(\xi)$ для любого $\alpha(\xi)$ в (6.8), получаем

$$\begin{aligned}
&\mathbb{E}_{\xi,\mathbf{e}} \left[\|\mathbf{e}\|_q^{1+\kappa} |f(x+\tau\mathbf{e},\xi) - f(x-\tau\mathbf{e},\xi)|^{1+\kappa}\right] \\
&\leq \mathbb{E}_{\xi,\mathbf{e}} \left[\|\mathbf{e}\|_q^{1+\kappa} |(f(x+\tau\mathbf{e},\xi) - \alpha) - (f(x-\tau\mathbf{e},\xi) - \alpha)|^{1+\kappa}\right] \\
&\overset{(6.1)}{\leq} 2^\kappa \mathbb{E}_{\xi,\mathbf{e}} \left[\|\mathbf{e}\|_q^{1+\kappa} |f(x+\tau\mathbf{e},\xi) - \alpha|^{1+\kappa}\right] \\
&+ 2^\kappa \mathbb{E}_{\xi,\mathbf{e}} \left[\|\mathbf{e}\|_q^{1+\kappa} |f(x-\tau\mathbf{e},\xi) - \alpha|^{1+\kappa}\right]. \qquad (6.10)
\end{aligned}$$

Мы учитываем, что распределение $\mathbf{e}$ симметрично,

$$(6.10) \leq 2^{\kappa+1} \mathbb{E}_{\xi,\mathbf{e}} \left[\|\mathbf{e}\|_q^{1+\kappa} |f(x+\tau\mathbf{e},\xi) - \alpha|^{1+\kappa}\right]. \qquad (6.11)$$

Пусть $\alpha(\xi) = \mathbb{E}_{\mathbf{e}}[f(x+\tau\mathbf{e},\xi)]$, тогда из-за неравенства Коши-Буняковского и свойств условного ожидания, получаем

$$\begin{aligned}
(6.11) \ &\leq \ 2^{\kappa+1}\mathbb{E}_{\xi,\mathbf{e}}\left[\|\mathbf{e}\|_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi)-\alpha|^{1+\kappa}\right] \\
&= \ 2^{\kappa+1}\mathbb{E}_{\xi}\left[\mathbb{E}_{\mathbf{e}}\left[\|\mathbf{e}\|_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi)-\alpha|^{1+\kappa}\right]\right] \\
&\leq \ 2^{\kappa+1}\sqrt{\mathbb{E}_{\mathbf{e}}\left[\|\mathbf{e}\|_q^{2(1+\kappa)}\right]}\cdot \\
&\cdot \ \mathbb{E}_{\xi}\left[\sqrt{\mathbb{E}_{\mathbf{e}}\left[|f(x+\tau\mathbf{e},\xi)-\mathbb{E}_{\mathbf{e}}[f(x+\tau\mathbf{e},\xi)]|^{2(1+\kappa)}\right]}\right]. \qquad (6.12)
\end{aligned}$$

Далее мы используем неравенство $\mathbb{E}_{\mathbf{e}}\left[\|\mathbf{e}\|_q^{2(1+\kappa)}\right] \leq a_q^{2(1+\kappa)}$ и Лемму 6.1.3 для функции $f(x+\tau\mathbf{e},\xi)$ с фиксированным ξ и константой Липшеца $M_2(\xi)\tau$,

$$\begin{aligned}
(6.12) \ &\leq \ 2^{\kappa+1}a_q^{1+\kappa}\mathbb{E}_{\xi}\left[\sqrt{\left(\frac{2^{-1/2}\tau^2M_2^2(\xi)}{d}\right)^{1+\kappa}}\right] \\
&= \ 2^{\kappa+1}a_q^{1+\kappa}\left(\frac{\tau^2 2^{-1/2}}{d}\right)^{(1+\kappa)/2}\mathbb{E}_{\xi}\left[M_2^{1+\kappa}(\xi)\right] \\
&\leq \ 2^{\kappa+1}\left(\sqrt{\frac{2^{-1/2}}{d}}a_qM_2\tau\right)^{1+\kappa}. \qquad (6.13)
\end{aligned}$$

Далее мы разребёмся со слагаемым (6.9). Мы применим неравенство Коши-Буняковского, Предположение 3 об ограничении шума и неравенство $\mathbb{E}_{\mathbf{e}}\left[\|\mathbf{e}\|_q^{2(1+\kappa)}\right] \leq a_q^{2(1+\kappa)}$, которое следует из определения a_q. В итоге мы получим

$$\begin{aligned}
&\mathbb{E}_{\xi,\mathbf{e}}\left[\|\mathbf{e}\|_q^{1+\kappa}|\delta(x+\tau\mathbf{e})-\delta(x-\tau\mathbf{e})|^{1+\kappa}\right] \\
&\leq \sqrt{\mathbb{E}_{\mathbf{e}}\left[\|\mathbf{e}\|_q^{2(1+\kappa)}\right]\mathbb{E}_{\mathbf{e}}\left[|\delta(x+\tau\mathbf{e})-\delta(x-\tau\mathbf{e})|^{2(1+\kappa)}\right]} \\
&\leq a_q^{1+\kappa}2^{1+\kappa}\Delta^{1+\kappa} = (2a_q\Delta)^{1+\kappa}. \qquad (6.14)
\end{aligned}$$

Сложив (6.13) и(6.14), мы доказываем финальное неравенство

$$\mathbb{E}_{\xi,\mathbf{e}}[\|g(x,\xi,\mathbf{e})\|_q^{1+\kappa}] \leq \frac{1}{2}\left(\frac{d}{\tau}\right)^{1+\kappa}\left(2^{1+\kappa}\left(\sqrt{\frac{2^{-1/2}}{d}}a_q\tau M_2\right)^{1+\kappa} + (2a_q\Delta)^{1+\kappa}\right) =$$

$$= 2^\kappa\left(\frac{\sqrt{d}}{2^{1/4}}a_q M_2\right)^{1+\kappa} + 2^\kappa\left(\frac{d a_q\Delta}{\tau}\right)^{1+\kappa}.$$

□

6.2 Доказательство сходимости по мат. ожиданию Алгоритма с Устойчивым SMD

Теорема 6.2.1. *Пусть функция f, удовлетворяющая Предположениям 1, 2, 3, $q \in [1+\kappa, \infty]$, количество итераций T, константа сглаживания $\tau > 0$ заданы заранее. Выберем $\left(1, \frac{1+\kappa}{\kappa}\right)$-равномерно выпуклую по ℓ_p-норме прокс-функцию $\Psi_p(x)$ (К примеру, $\Psi_p(x) = K_q^{1/\kappa}\phi_p(x)$, где K_q, ϕ_p определены в (2.6) и (2.7) соответственно). Установим размер шага равный $\nu = \frac{R_0^{1/\kappa}}{\sigma_q}T^{-\frac{1}{1+\kappa}}$ с σ_q из Леммы 2.0.2, расстоянием от начальной точки x_0 до решения x^* $R_0^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}D_{\Psi_p}(x^*, x_0)$ и диаметром компакта $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}\sup\limits_{x,y\in\mathcal{X}} D_{\Psi_p}(x,y)$. Пусть $\overline{x}_T$ результат работы Алгоритма 1 с заданными выше параметрами.*

1. *Тогда имеем следующую оценку*

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_0\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \tag{6.15}$$

где $\sigma_q^{1+\kappa} = 2^\kappa\left(\frac{\sqrt{d}}{2^{1/4}}a_q M_2\right)^{1+\kappa} + 2^\kappa\left(\frac{d a_q\Delta}{\tau}\right)^{1+\kappa}$.

2. К тому же, с оптимальным $\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4R_0 d a_q \Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$*, имеем*

$$\begin{aligned}\mathbb{E}[f(\overline{x}_T)] - f(x^*) \;&\leq\; \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_0da_q\Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} \\ &+\; \frac{2\sqrt{d}a_qM_2R_0}{T^{\frac{\kappa}{1+\kappa}}}. \end{aligned} \tag{6.16}$$

Доказательство. По определению $x_* \in \arg\min_{x\in\mathcal{X}} f(x)$.

Мы используем Теорему Сходимости 2.0.4 для Устойчивого SMD и зафиксируем векторы обновлений $g_k(x_k, \xi_k, \mathbf{e}_k)$

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle g_{k+1}, x_k - x^*\rangle \leq \frac{\kappa}{\kappa+1}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\|g_{k+1}\|_q^{1+\kappa}. \tag{6.17}$$

Затем мы берем полное математическое ожидание $\mathbb{E}$ с обеих сторон (6.17)

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\langle g_{k+1}, x_k - x^*\rangle\right] \leq \frac{\kappa}{\kappa+1}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|g_{k+1}\|_q^{1+\kappa}\right]. \tag{6.18}$$

Используя ограниченность $(1+\kappa)$-го момента оценки градиента из Леммы 6.1.7 для правой части неравенства (6.18), получаем

$$\frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|g_{k+1}\|_q^{1+\kappa}\right] \leq \frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\sigma_q^{1+\kappa} \leq \frac{\nu^\kappa}{1+\kappa}\sigma_q^{1+\kappa}. \tag{6.19}$$

Используя условное математическое ожидание и Лемму 6.1.6 для левой части неравенства (6.18), мы оцениваем

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\langle g_{k+1}, x_k - x^*\rangle\right] = \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\mathbb{E}_{|\leq k}[\langle g_{k+1}, x_k - x^*\rangle]\right]$$

$$\geq \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle] - \frac{1}{T}\sum_{k=0}^{T-1}\frac{d\Delta}{\tau}\mathbb{E}\left[\mathbb{E}_{\mathbf{e}_k|\leq k}[|\langle \mathbf{e}_k, x_k - x^*\rangle|]\right]. \quad (6.20)$$

1. Для первого слагаемого (6.20) в силу выпуклости $\hat{f}_\tau(x)$ получаем

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle] \geq \frac{1}{T}\sum_{k=0}^{T-1}\Big(\mathbb{E}[\hat{f}_\tau(x_k)] - \hat{f}_\tau(x_*)\Big).$$

Затем мы определяем $\overline{x}_T = \frac{1}{T}\sum\limits_{k=0}^{T-1} x_k$ и используем неравенство Йенсена

$$\frac{1}{T}\sum_{k=0}^{T-1}\Big(\mathbb{E}[\hat{f}_\tau(x_k)] - \hat{f}_\tau(x_*)\Big) \geq \mathbb{E}[\hat{f}_\tau(\overline{x}_T)] - \hat{f}_\tau(x^*).$$

Наконец, мы применяем свойство аппроксимации $\hat{f}_\tau(x)$ из Леммы 2.0.1

$$\mathbb{E}[\hat{f}_\tau(\overline{x}_T)] - \hat{f}_\tau(x^*) \geq \mathbb{E}[f(\overline{x}_T)] - f(x^*) - 2M_2\tau. \quad (6.21)$$

2. Для второго слагаемого (6.20) мы используем свойство концентрации меры из Леммы 6.1.5 и оцениваем

$$\begin{aligned}
&-\frac{d\Delta}{T\tau}\sum_{k=0}^{T-1}\mathbb{E}_{\mathbf{e}_k|\leq k}[|\langle \mathbf{e}_k, x_k - x^*\rangle|] \\
&\qquad \geq -\frac{d\Delta}{T\tau}\sum_{k=0}^{T-1}\frac{1}{\sqrt{d}}\|x_k - x^*\|_2 \\
&\qquad \overset{p\leq 2}{\geq} -\frac{d\Delta}{T\tau}\sum_{k=0}^{T-1}\frac{1}{\sqrt{d}}\|x_k - x^*\|_p. \quad (6.22)
\end{aligned}$$

Заметим, что Ψ_p является $\big(1, \frac{1+\kappa}{\kappa}\big)$-равномерно выпуклой функцией по

p норме. Тогда по определению (2.5) мы оцениваем $\|x_k - x^*\|_p$

$$\|x_k - x^*\|_p \leq \left(\frac{1+\kappa}{\kappa} D_{\Psi_p}(x_k, x^*)\right)^{\frac{\kappa}{1+\kappa}} \leq \sup_{x,y\in\mathcal{X}} \left(\frac{1+\kappa}{\kappa} D_{\Psi_{q^*}}(x, y)\right)^{\frac{\kappa}{1+\kappa}} = \mathcal{D}_\Psi$$

Следовательно, после этой оценки мы получаем

$$(6.22) \geq -\frac{d\Delta}{T\tau} \sum_{k=0}^{T-1} \frac{1}{\sqrt{d}} \|x_k - x^*\|_p \geq -\frac{\sqrt{d}\Delta}{\tau} \mathcal{D}_\Psi. \tag{6.23}$$

Далее мы объединяем (6.19), (6.21), (6.23) вместе и получаем итоговую границу

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau} \mathcal{D}_\Psi + \frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa} \sigma_q^{1+\kappa}. \tag{6.24}$$

Теперь мы выбираем хорошие параметры Алгоритма, чтобы минимизировать правую часть (6.24). Выбрав оптимальный $\nu = \frac{R_0^{1/\kappa}}{\sigma_q} T^{-\frac{1}{1+\kappa}}$, мы снизим границу

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau} \mathcal{D}_\Psi + 2R_0\sigma_q T^{-\frac{\kappa}{1+\kappa}}.$$

Наконец, мы получаем явную оценку σ_q, используя Лемму 6.1.1

$$\sigma_q \leq 2\left(\frac{\sqrt{d}}{2^{1/4}} a_q M_2\right) + 2\left(\frac{d a_q \Delta}{\tau}\right),$$

И выбираем оптимальный τ

$$\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4R_0 d a_q \Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}.$$

□

6.3 Доказательство сходимости по мат. ожиданию Алгоритма с Клиппингом

Сначала мы докажем несколько полезных утверждений о свойствах клиппированного вектора градиента. Аналогичное доказательство можно найти в работе [33].

Лемма 6.3.1. *Для $c > 0$ и случайного вектора $g = g(x,\xi,\mathbf{e})$ мы определяем $\hat{g} = \frac{g}{\|g\|_q}\min(\|g\|_q, c)$. Тогда имеем*

1.
$$\|\hat{g} - \mathbb{E}[\hat{g}]\|_q \leq 2c. \tag{6.25}$$

2. *Также если* $\mathbb{E}[\|g(x,\xi,\mathbf{e})\|_q^{1+\kappa}] \leq \sigma_q^{1+\kappa}$, *то дополнительно верно*

   (a)
$$\mathbb{E}[\|\hat{g}\|_q^2] \leq \sigma_q^{1+\kappa}c^{1-\kappa}, \tag{6.26}$$

   (b)
$$\mathbb{E}[\|\hat{g} - \mathbb{E}[\hat{g}]\|_q^2] \leq 4\sigma_q^{1+\kappa}c^{1-\kappa}, \tag{6.27}$$

   (c)
$$\|\mathbb{E}[g] - \mathbb{E}[\hat{g}]\|_q \leq \frac{\sigma_q^{1+\kappa}}{c^\kappa}. \tag{6.28}$$

Доказательство. 1. По неравенству Йенсена для $\|\cdot\|_q$ и определению $\hat{g}$ оцениваем

$$\begin{aligned}
\|\hat{g} - \mathbb{E}[\hat{g}]\|_q &\leq \|\hat{g}\|_q + \|\mathbb{E}[\hat{g}]\|_q \\
&\leq \left\|\frac{g}{\|g\|_q}\min(\|g\|_q, c)\right\|_q + \mathbb{E}\left[\left\|\frac{g}{\|g\|_q}\min(\|g\|_q, c)\right\|_q\right] \\
&= \min(\|g\|_q, c) + \mathbb{E}[\min(\|g\|_q, c)] \\
&\leq c + c = 2c.
\end{aligned} \tag{6.29}$$

2. (a) Учитывая $\mathbb{E}[\|g(x,\xi,\mathbf{e})\|_q^{1+\kappa}] \le \sigma_q^{1+\kappa}$ и $\|\hat{g}\|_q \le c$, мы получаем

$$\mathbb{E}[\|\hat{g}\|_q^{1+\kappa}\|\hat{g}\|_q^{1-\kappa}] \le \sigma_q^{1+\kappa}c^{1-\kappa}.$$

(b) По неравенству Йенсена для $\|\cdot\|_q$ получаем

$$\begin{aligned}\mathbb{E}[\|\hat{g}-\mathbb{E}[\hat{g}]\|_q^2] &\le 2\mathbb{E}[\|\hat{g}\|_q^2 + 2\|\mathbb{E}[\hat{g}]\|_q^2] \\ &\le 2\mathbb{E}[\|\hat{g}\|_q^2] + 2\mathbb{E}[\|\hat{g}\|_q^2]] \\ &\overset{(6.26)}{\le} 2\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa} + 2\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa} \\ &\le 4\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa}. \end{aligned} \tag{6.30}$$

(c) В силу выпуклости функции нормы и неравенства Йенсена оцениваем

$$\|\mathbb{E}[g]-\mathbb{E}[\hat{g}]\|_q \le \mathbb{E}[\|g-\hat{g}\|_q] \le \mathbb{E}\left[\|g\|_q \mathbb{1}_{\{\|g\|_q>c\}}\right].$$

Окончательный результат следует из

$$\|g\|_q^{1+\kappa}\mathbb{1}_{\{\|g\|_q>c\}} \ge \|g\|_q c^{\kappa}\mathbb{1}_{\{\|g\|_q>c\}},$$

т.е.

$$\mathbb{E}\left[\|g\|_q \mathbb{1}_{\{\|g\|_q>c\}}\right] \le \mathbb{E}\left[\|g\|_q \mathbb{1}_{\{\|g\|_q>c\}}\right] \le \frac{\sigma_{q,\kappa}^{1+\kappa}}{c^{\kappa}}. \tag{6.31}$$

□

Теорема 6.3.2. *Пусть функция f, удовлетворяющая Предположениям 1, 2, 3, $q \in [2,\infty]$, количество итераций T, константа сглаживания $\tau > 0$ заданы заранее. Выберем 1-сильно выпуклую функцию по p-норме прокс-функцию $\Psi_p(x)$. Установим размер шага $\nu = \left(\frac{R_0^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}}\right)^{\frac{1}{1+\kappa}}$ с σ_q из Леммы 2.0.2, расстоянием от начальной точки x_0 до решения x^* $R_0^{\frac{1+\kappa}{\kappa}} \overset{def}{=}$*

$\frac{1+\kappa}{\kappa} D_{\Psi_p}(x^*, x_0)$ *и диаметром компакта* $\mathcal{D}_{\Psi}^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa} \sup_{x,y \in \mathcal{X}} D_{\Psi_p}(x,y)$. *После установим константу клиппирования* $c = \frac{2\kappa \mathcal{D}_\Psi}{(1-\kappa)\nu}$. *Пусть* $\overline{x}_T$ *результат работы Алгоритма 2 с заданными выше параметрами.*

1. *Тогда имеем следующую оценку*

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_0^{\frac{2\kappa}{1+\kappa}} \mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}} \sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \tag{6.32}$$

где $\sigma_q^{1+\kappa} = 2^\kappa \left(\frac{\sqrt{d}}{2^{1/4}} a_q M_2\right)^{1+\kappa} + 2^\kappa \left(\frac{d a_q \Delta}{\tau}\right)^{1+\kappa}$.

2. *К тому же, с оптимальным* $\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4R_0^{\frac{2\kappa}{1+\kappa}} \mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}} d a_q \Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$, *имеем*

$$\begin{aligned} \mathbb{E}[f(\overline{x}_T)] - f(x^*) &\leq \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32 M_2 R_0^{\frac{2\kappa}{1+\kappa}} \mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}} d a_q \Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} \\ &+ \frac{2\sqrt{d} a_q M_2 R_0^{\frac{2\kappa}{1+\kappa}} \mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}}{T^{\frac{\kappa}{1+\kappa}}}. \end{aligned} \tag{6.33}$$

Доказательство. Заметим из первого слагаемого (6.20) в доказательстве Теоремы 3.0.1, что для любого x_k

$$f(\overline{x}_T) - f(x^*) \leq \frac{1}{T} \sum_{k=0}^{T-1} \langle \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle + 2M_2\tau. \tag{6.34}$$

Далее мы определяем функции

$$l_k(x) \overset{\text{def}}{=} \langle \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x - x^* \rangle.$$

Заметим, что $l_k(x)$ выпукла для любого k и $\nabla l_k(x) = \mathbb{E}_{|\leq k}[\hat{g}_{k+1}]$. Следовательно, $\hat{g_{k+1}}$ является несмещенной оценкой градиента $l_k(x)$. С помощью этих функций мы можем переписать правую часть (6.34) как

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle + 2M_2\tau$$

$$= \frac{1}{T}\sum_{k=0}^{T-1}\Big(\langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle\Big) + \frac{1}{T}\sum_{k=0}^{T-1}\left(l_k(x_k) - l_k(x^*)\right) + 2M_2\tau. \tag{6.35}$$

Затем мы берем полное математическое ожидание с обеих сторон (6.35)

$$\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle\right] + 2M_2\tau$$

$$\begin{aligned} = \ &\underbrace{\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\Big(\langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle\Big)\right]}_{\text{D}} \\ + \ &\underbrace{\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\left(l_k(x_k) - l_k(x^*)\right)\right]}_{\text{E}} + 2M_2\tau. \end{aligned} \tag{6.36}$$

Мы добавляем $\pm\mathbb{E}_{|\leq k}[g_{k+1}]$ к слагаемому D и получаем

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\Big(\langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle\Big)\right] \\ &= \mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\langle \mathbb{E}_{|\leq k}[g_{k+1}] - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle\right] \\ &+ \mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[g_{k+1}], x_k - x^*\rangle\right]. \end{aligned} \tag{6.37}$$

Чтобы оценить первый член (6.37), заметим, что Ψ_p является $(1,2)$-равномерно выпуклой функцией по p норме. Тогда по определению (2.5) мы ограничим $\|x_k - x^*\|_p$ как

$$\|x_k - x^*\|_p \le \left(2D_{\Psi_p}(x_k, x^*)\right)^{\frac{1}{2}} \le \sup_{x,y\in\mathcal{X}} \left(2D_{\Psi_p}(x, y)\right)^{\frac{1}{2}} = \mathcal{D}_\Psi,$$

и оценим $\|x_k - u\|_p \le \mathcal{D}_\Psi, \forall u \in \mathcal{X}$.

Применим неравенство Коши–Буняковского к скалярному произведению в первом члене уравнения (6.37)

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\left(\langle \mathbb{E}_{|\le k}[g_{k+1}] - \mathbb{E}_{|\le k}[\hat{g}_{k+1}], x_k - x^*\rangle\right)\right] \\
&\le \frac{1}{T}\sum_{k=0}^{T-1}\left(\mathbb{E}\left[\mathbb{E}_{|\le k}\left[\|\mathbb{E}_{|\le k}[g_{k+1}] - \mathbb{E}_{|\le k}[\hat{g}_{k+1}]\|_q \|x_k - x^*\|_p\right]\right]\right) \overset{(6.28)}{\le} \mathcal{D}_\Psi \frac{\sigma_q^{1+\kappa}}{c^\kappa}.
\end{aligned} \tag{6.38}$$

Чтобы ограничить второй член в (6.37) мы используем Лемму 6.1.6 и Лемму 6.1.5

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\left(\langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\le k}[g_{k+1}], x_k - x^*\rangle\right)\right] \\
&\le \frac{1}{T}\sum_{k=0}^{T-1}\frac{d\Delta}{\tau}\mathbb{E}\left[\mathbb{E}_{\mathbf{e}|<k}[|\langle \mathbf{e}, x_k - x^*\rangle|]\right] \\
&\le \frac{1}{T}\sum_{k=0}^{T-1}\frac{d\Delta}{\tau}\frac{1}{\sqrt{d}}\mathbb{E}[\|x_k - x^*\|_2]
\end{aligned}$$

$$\overset{p\le 2}{\le} \frac{1}{T}\sum_{k=0}^{T-1}\frac{d\Delta}{\tau}\frac{1}{\sqrt{d}}\mathbb{E}[\|x_k - x^*\|_p] \le \frac{\Delta\sqrt{d}}{\tau}\mathcal{D}_\Psi. \tag{6.39}$$

Далее мы ограничиваем слагаемое E. Прежде всего, мы перепишем его как

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}(l_k(x_k)-l_k(x^*))\right] &= \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\mathbb{E}_{|\leq k}[\langle\mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k-x^*\rangle]\right]\\ &= \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\mathbb{E}_{|\leq k}[\langle\hat{g}_{k+1}, x_k-x^*\rangle]\right].\end{aligned}$$

Для Устойчивого SMD с векторами обновления $\hat{g}_k$ по Теореме о Сходимости 2.0.4 с ограниченным вторым моментом верно следующее неравенство

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle\hat{g}_{k+1}, x_k-x^*\rangle \leq \frac{1}{2}\frac{R_0^2}{\nu T}+\frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}\|\hat{g}_{k+1}\|_q^2. \tag{6.40}$$

Взяв $\mathbb{E}$ с обеих сторон (6.40), получим

$$\begin{aligned}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\langle\hat{g}_{k+1}, x_k-x^*\rangle] &= \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\mathbb{E}_{|\leq k}[\langle\hat{g}_{k+1}, x_k-x^*\rangle]\right]\\ &\leq \frac{1}{2}\frac{R_0^2}{\nu T}+\frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\mathbb{E}_{|\leq k}[\|\hat{g}_{k+1}\|_q^2]\right]. \quad (6.41)\end{aligned}$$

По (6.26) из Леммы 6.3.1 мы ограничиваем второй момент клиппированного градиента

$$\mathbb{E}_{|\leq k}(\|\hat{g}_{k+1}\|_q^2) \leq \sigma_q^{1+\kappa}c^{1-\kappa},$$

И, следовательно, получаем

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\langle\hat{g}_{k+1}, x_k-x^*\rangle] \leq \frac{1}{2}\frac{R_0^2}{\nu T}+\frac{\nu}{2}\sigma_q^{1+\kappa}c^{1-\kappa}. \tag{6.42}$$

Объединив слагаемые (6.38), (6.39), (6.42), мы вычислим итоговую границу

$$\mathbb{E}[f(\overline{x}_T)]-f(x^*) \leq 2M_2\tau+\frac{1}{2}\frac{R_0^2}{\nu T}+\frac{\nu}{2}\sigma_q^{1+\kappa}c^{1-\kappa}+\left(\frac{\sigma_q^{1+\kappa}}{c^\kappa}+\Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi.$$

Для получения минимальной верхней границы находим оптимальные параметры. Во-первых, мы выбираем c, находя минимум

$$\min_{c>0} \sigma_q^{1+\kappa} \left(\frac{1}{c^\kappa} \mathcal{D}_\Psi + \frac{\nu}{2} c^{1-\kappa} \right) = \min_c \sigma_q^{1+\kappa} h_1(c)$$

$$h_1'(c) = \tfrac{\nu}{2}(1-\kappa)c^{-\kappa} - \kappa \tfrac{1}{c^{1+\kappa}} \mathcal{D}_\Psi = 0 \Rightarrow c^* = \tfrac{2\kappa \mathcal{D}_\Psi}{(1-\kappa)\nu}.$$

$$\begin{aligned} \mathbb{E}[f(\overline{x}_T)] - f(x^*) &\le 2M_2\tau + \frac{1}{2}\frac{R_0^2}{\nu T} + \Delta \frac{\sqrt{d}}{\tau} \mathcal{D}_\Psi \\ &+ \sigma_{q,\kappa}^{1+\kappa} \left(\mathcal{D}^{1-\kappa} 2^{-\kappa} \nu^\kappa \left[\frac{(1-\kappa)^\kappa}{\kappa^\kappa} + \frac{\kappa^{(1-\kappa)}}{(1-\kappa)^{(1-\kappa)}} \right] \right) \end{aligned} \tag{6.43}$$

Учитывая оценку $\kappa \in [0,1]$ и, как следствие,

$$\left[\frac{(1-\kappa)^\kappa}{\kappa^\kappa} + \frac{\kappa^{(1-\kappa)}}{(1-\kappa)^{(1-\kappa)}} \right] \le 2,$$

мы упрощаем оценку (6.43)

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \frac{1}{2}\frac{R_0^2}{\nu T} + \Delta \frac{\sqrt{d}}{\tau} \mathcal{D}_\Psi + \sigma_q^{1+\kappa} \left(2\mathcal{D}_\Psi^{1-\kappa} \nu^\kappa \right). \tag{6.44}$$

Аналогично выбирая оптимальные ν^*, получаем

$$\nu^* = \left(\frac{R_0^2}{4T\kappa \sigma_q^{1+\kappa} \mathcal{D}_\Psi^{1-\kappa}} \right)^{\frac{1}{1+\kappa}}$$

И

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \Delta \frac{\sqrt{d}}{\tau} \mathcal{D}_\Psi + \frac{R_0^{\frac{2\kappa}{1+\kappa}} \mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}} \sigma_q}{T^{\frac{\kappa}{1+\kappa}}} 2 \left[\kappa^{\frac{1}{1+\kappa}} + \kappa^{-\frac{\kappa}{1+\kappa}} \right].$$

С учетом границы $\kappa \in [0,1]$ верно следующее неравенство

$$\left[\kappa^{\frac{1}{1+\kappa}} + \kappa^{-\frac{\kappa}{1+\kappa}} \right] \le 2.$$

Таким образом, мы можем еще больше упростить верхнюю границу

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \Delta\frac{\sqrt{d}}{\tau}\mathcal{D}_\Psi + 2\frac{R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}. \tag{6.45}$$

Чтобы избежать $\nu \to \infty$ при $\kappa \to 0$, можно также выбрать $\nu^* = \left(\frac{R_0^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}}\right)^{\frac{1}{1+\kappa}}$. Оценка (6.45) не изменится.

Наконец, мы получаем явную оценку σ_q с помощью Леммы 6.1.1

$$\sigma_q \leq 2\left(\frac{\sqrt{d}}{2^{1/4}}a_q M_2\right) + 2\left(\frac{da_q\Delta}{\tau}\right),$$

И выбираем оптимальный τ

$$\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}da_q\Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}.$$

□

6.4 Доказательство сходимости с высокой вероятностью Алгоритма с Клиппингом

Для следующего доказательства нам понадобятся некоторые классические результаты концентрации меры. Неравенство Бернштейна для суммы мартингальных разностей (Лемма 23 из [33]).

Лемма 6.4.1. *Пусть* $\{X_i\}_{i\geq 1}$ *являются последовательностью мартингальных разностей, т.е.* $\mathbb{E}[X_i|X_{i-1},\ldots,X_1] = 0, \forall i \geq 1$. *При этом* b, σ *такие константы, что* $|X_i| < b$ *и* $\mathbb{E}[X_i^2|X_{i-1},\ldots,X_1] < \sigma^2$ *почти наверняка для* $i \geq 1$. *Тогда для произвольного фиксированного числа* μ *и для всех* T

с вероятностью не менее $1-\delta$ справедливо следующее неравенство

$$\left|\sum_{i=1}^{t} \mu X_i\right| \leq 2b|\mu| \log\frac{1}{\delta} + \sigma|\mu|\sqrt{2T\log\frac{1}{\delta}}.$$

Аналогичная лемма для суммы квадратов ограниченных случайных величин представлена в Теореме 20 из [33].

Лемма 6.4.2. *Пусть Z_i — последовательность случайных величин, адаптированная к фильтрации $\mathcal{F}_t$. Далее предположим, что $|Z_i| < b, \mathbb{E}[Z_i^2] \leq \sigma^2$ почти наверняка. Тогда для любого $\mu > 0$ с вероятностью не менее $1-\delta$ верно следующее неравенство*

$$\begin{aligned}\sum_{k=1}^{T} Z_k^2 \ \leq \ & 3T\sigma^2 \log\left(\frac{4}{\delta}\left[\log\left(\sqrt{\frac{\sigma^2 T}{\mu^2}}\right)+2\right]^2\right) \\ + \ & 20\max(\mu^2, b^2)\log\left(\frac{112}{\delta}\left[\log\left(\frac{2\max(\mu,b)}{\mu}\right)+1\right]^2\right). \end{aligned} \tag{6.46}$$

Выбрав $\mu = b \geq \sigma$, граница упрощается как

$$\sum_{k=1}^{T} Z_k^2 \leq 3T\sigma^2 \log\left(\frac{4}{\delta}\left[\log\left(\sqrt{T}\right)+2\right]^2\right) + 20b^2\log\left(\frac{12}{\delta}\right).$$

Теорема 6.4.3. *Пусть функция f, удовлетворяющая Предположениям 1, 2, 3, $q \in [2,\infty]$, количество итераций T, константа сглаживания $\tau > 0$ заданы заранее. Выберем 1-сильно выпуклую функцию по p-норме прокс-функцию $\Psi_p(x)$. Установим константу клиппирования равную $c = T^{\frac{1}{(1+\kappa)}}\sigma_q$ с σ_q из Леммы 2.0.2. Установим размер шага $\nu = \frac{\mathcal{D}_\Psi}{c}$ с диаметром компакта $\mathcal{D}_\Psi^2 \overset{def}{=} 2\sup_{x,y\in\mathcal{X}} D_{\Psi_p}(x,y)$. Пусть $\overline{x}_T$ результат работы Алгоритма 2 с заданными выше параметрами. Помимо этого для $\delta \in [0,1)$ мы обозначим $\tilde{\delta}^{-1} = \frac{4}{\delta}\left[\log\left(\sqrt{T}\right)+2\right]^2$ и $\beta = \left[3 + 8\log\frac{1}{\delta} + 12\log\frac{1}{\tilde{\delta}} + 20\log\frac{4}{\delta} \ + 4\sqrt{2\log\frac{1}{\delta}}\right]$.*

1. *Тогда с вероятностью не менее $1-\delta$ мы имеем оценку*

$$f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \frac{\Delta\sqrt{d}}{\tau}\mathcal{D}_\Psi + \frac{\mathcal{D}_\Psi \sigma_q \beta}{2T^{\frac{\kappa}{1+\kappa}}}, \tag{6.47}$$

где $\sigma_q^{1+\kappa} = 2^\kappa \left(\frac{\sqrt{d}}{2^{1/4}} a_q M_2\right)^{1+\kappa} + 2^\kappa \left(\frac{d a_q \Delta}{\tau}\right)^{1+\kappa}$.

2. *К тому же, с оптимальным* $\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 2\beta\mathcal{D}_\Psi d a_q \Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$, *имеем*

$$f(\overline{x}_T) - f(x^*) \leq \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + 4\sqrt{\frac{\beta M_2 \mathcal{D}_\Psi d a_q \Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{\beta\sqrt{d} a_q M_2 \mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}. \tag{6.48}$$

Доказательство. Заметим по первому слагаемому (6.20) в доказательстве Теоремы 3.0.1, что для любого x_k верно следующее неравенство

$$f(\overline{x}_T) - f(x^*) \leq \frac{1}{T}\sum_{k=0}^{T-1} \langle \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle + 2M_2\tau. \tag{6.49}$$

Для Устойчивого SMD с векторами обновления $\hat{g}_k$ Теорема сходимости 2.0.4 с ограниченным вторым моментом гарантирует, что

$$\frac{1}{T}\sum_{k=0}^{T-1} \langle \hat{g}_{k+1}, x_k - x^* \rangle \leq \frac{1}{2}\frac{R_0^2}{\nu T} + \frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1} \|\hat{g}_{k+1}\|_q^2. \tag{6.50}$$

Определим случайную величину $Z_k = \|\hat{g}_{k+1}\|_q$ и заметим, что $|Z_k| \leq c$ по определению клиппинга и $\mathbb{E}[Z_i^2] \leq 4\sigma_{q,\kappa}^{1+\kappa} c^{1-\kappa}$ по (6.27) из свойств клиппированного градиента из Леммы 6.3.1. Таким образом, мы можем применить лемму 6.4.2 и с вероятностью не менее $1-\delta$ ограничить среднюю сумму вторых моментов клиппированных градиентов

$$\frac{1}{T}\sum_{k=0}^{T-1} \|\hat{g}_{k+1}\|_q^2 \leq 12\sigma_{q,\kappa}^{1+\kappa} c^{1-\kappa} \log\left(\frac{4}{\delta}\left[\log\left(\sqrt{T}\right) + 2\right]^2\right) + \frac{20}{T}c^2 \log\left(\frac{12}{\delta}\right). \tag{6.51}$$

Левую часть (6.50) можно переписать как

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1}, x_k - x^*\rangle = \frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1} - \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle + \frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle$$

$$= \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1} - \mathbb{E}_{|\le k}[\hat{g}_{k+1}], x_k - x^*\rangle}_{1} + \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}\langle \mathbb{E}_{|\le k}[\hat{g}_{k+1}] - \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle}_{2}$$

$$+ \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle}_{3}.$$

Для слагаемого 1 мы можем доказать, что это сумма последовательности мартингальных разностей. Действительно, заметим, что при фиксированном x_k, когда мы берем $\mathbb{E}_{|\le k}$, выполняется свойство мартингала, т.е.

$$\mathbb{E}_{|\le k}[\langle \hat{g}_{k+1} - \mathbb{E}_{|\le k}[\hat{g}_{k+1}], x_k - x^*\rangle] = 0.$$

По (6.25) из Леммы 6.3.1 мы ограничиваем каждый элемент последовательности мартингальных разностей

$$|\langle \hat{g}_{k+1} - \mathbb{E}_{|\le k}[\hat{g}_{k+1}], x_k - x^*\rangle| \le \|\hat{g}_{k+1} - \mathbb{E}_{|\le k}[\hat{g}_{k+1}]\|_q \|x_k - x^*\|_p \le 2c \cdot \|x_k - x^*\|_p.$$

Также по неравенству (6.27) из леммы 6.3.1 мы ограничиваем математическое ожидание квадрата каждого элемента

$$\mathbb{E}\left[|\langle \hat{g}_{k+1} - \mathbb{E}_{|\le k}[\hat{g}_{k+1}], x_k - x^*\rangle|^2\right] \le 4\sigma_q^{1+\kappa} c^{1-\kappa} \cdot \|x_k - x^*\|_p^2.$$

Заметим, что Ψ_p является $(1,2)$-равномерно выпуклой функцией по p норме. Тогда по определению (2.5) мы оцениваем

$$\|x_k - x^*\|_p \le \left(2D_{\Psi_p}(x_k, x^*)\right)^{\frac{1}{2}} \le \sup_{x,y\in\mathcal{X}} \left(2D_{\Psi_p}(x, y)\right)^{\frac{1}{2}} = \mathcal{D}_\Psi,$$

и ограничиваем $\|x_k - u\|_p \le \mathcal{D}, \forall u \in \mathcal{X}$. Следовательно, мы можем применить неравенство Бернштейна 6.4.1 с $\mu = \frac{1}{T}$ и получить с вероятностью не менее $1-\delta$, что

$$\frac{1}{T}\sum_{k=0}^{T-1} |\langle \hat{g}_{k+1} - \mathbb{E}_{|\le k}[\hat{g}_{k+1}], x_k - x^*\rangle| \le \frac{4c\mathcal{D}_\Psi}{T}\log\frac{1}{\delta} + \frac{\sqrt{4\sigma_q^{1+\kappa}c^{1-\kappa}}}{\sqrt{T}}\mathcal{D}_\Psi^2\sqrt{2\log\frac{1}{\delta}}. \tag{6.52}$$

Для 2 мы используем оценку слагаемого D из (6.36) из доказательства Теоремы 6.3.2

$$|\langle \mathbb{E}_{|\le k}[\hat{g}_{k+1}] - \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle| \le \left(\frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi. \tag{6.53}$$

Для 3 используем уже полученную оценку (6.49)

$$f(\overline{x}_T) - f(x^*) - 2M_2\tau \le \frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle. \tag{6.54}$$

Подставляя (6.51), (6.52), (6.53), (6.54) в (6.50), получаем с вероятностью не менее $1-\delta$, что

$$\begin{aligned} f(\overline{x}_T) - f(x^*) \le 2M_2\tau + \left(\frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi + \frac{1}{2}\frac{R_0^2}{\nu T} \\ +\frac{\nu}{2}\left[12\sigma_q^{1+\kappa}c^{1-\kappa}\log\left(\frac{4}{\delta}\left[\log\left(\sqrt{T}\right)+2\right]^2\right)\right] \\ +\frac{\nu}{2}\frac{20}{T}c^2\log\left(\frac{12}{\delta}\right) + \frac{4c\mathcal{D}_\Psi}{T}\log\frac{1}{\delta} + \frac{\sqrt{4\sigma_q^{1+\kappa}c^{1-\kappa}}}{\sqrt{T}}\mathcal{D}_\Psi^2\sqrt{2\log\frac{1}{\delta}}. \end{aligned} \tag{6.55}$$

Далее мы выбираем оптимальные параметры, чтобы минимизировать верхнюю границу. Выбрав $c = T^{\frac{1}{(1+\kappa)}}\sigma_q$ и подставив её в (6.55), мы получим

$$f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \left(\frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} + \Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi + \frac{1}{2}\frac{R_0^2}{\nu T}$$

$$+\frac{\nu}{2}\left[12\sigma_q^2 T^{\frac{1-\kappa}{(1+\kappa)}}\log\left(\frac{4}{\delta}\left[\log\left(\sqrt{T}\right)+2\right]^2\right)\right]$$

$$+\frac{\nu}{2}\frac{20\sigma_q^2}{T^{\frac{\kappa-1}{1+\kappa}}}\log\left(\frac{12}{\delta}\right) + \frac{4\sigma_q\mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}\log\frac{1}{\delta} + \frac{2\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\mathcal{D}_\Psi\sqrt{2\log\frac{1}{\delta}}. \tag{6.56}$$

Для удобства определим $\tilde{\delta}^{-1} = \frac{4}{\delta}\left[\log\left(\sqrt{T}\right)+2\right]^2$, выберем $\nu = \frac{\mathcal{D}_\Psi}{c}$, подставим её в (6.56), тем самым получив оценку

$$f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \left(\frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} + \Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi + \frac{\mathcal{D}_\Psi\sigma_q}{2T^{\frac{\kappa}{1+\kappa}}}\left[1 + 12\log\frac{1}{\tilde{\delta}} + 20\log\frac{4}{\delta}\right]$$

$$+\frac{4\sigma_q\mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}\log\frac{1}{\delta} + \frac{2\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\mathcal{D}_\Psi\sqrt{2\log\frac{1}{\delta}}. \tag{6.57}$$

Упростив (6.57), мы показываем, что

$$\begin{aligned} f(\overline{x}_T) - f(x^*) \leq\ & 2M_2\tau + \Delta\frac{\sqrt{d}}{\tau}\mathcal{D}_\Psi \\ & + \frac{\mathcal{D}_\Psi\sigma_q}{2T^{\frac{\kappa}{1+\kappa}}}\left[3 + 8\log\frac{1}{\delta} + 12\log\frac{1}{\tilde{\delta}} + 20\log\frac{4}{\delta}\ + 4\sqrt{2\log\frac{1}{\delta}}\right]. \end{aligned}$$

Наконец, мы получаем явную оценку σ_q с помощью Леммы 6.1.1

$$\sigma_q \leq 2\left(\frac{\sqrt{d}}{2^{1/4}}a_q M_2\right) + 2\left(\frac{d a_q\Delta}{\tau}\right),$$

и выберем оптимальный τ

$$\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 2\beta\mathcal{D}_\Psi d a_q\Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}.$$

$\square$

6.5 Набросок доказательства сходимости Алгоритма с Рестартами

В этой главе представлены наброски доказательств Теорем 5.0.1, 5.0.2.

Доказательство. В этом доказательстве $\tilde{O}(\cdot)$ обозначает $\log d$ множитель.

Шаг 1: Сходимость по мат. ожиданию Алгоритма с Устойчивым SMD.

Теперь x_0 в Алгоритме 1 можно выбирать случайным образом.

Аналогично доказательству теоремы 3.0.1, но с $\nu = \frac{\mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{1}{1+\kappa}}}{\sigma_q} T^{-\frac{1}{1+\kappa}}$ и ограничением $R_0 \le \mathcal{D}_\Psi$ можно получить из (6.24) оценку

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + 2\mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{\kappa}{1+\kappa}} \sigma_q T^{-\frac{\kappa}{1+\kappa}}. \quad (6.58)$$

При обязательном условии $\Delta \le \frac{\sigma_q^2 \mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{\kappa}{1+\kappa}}}{M_2\sqrt{d}T^{\frac{2\kappa}{1+\kappa}}}$ выбрав $\tau = \frac{\sigma_q \mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{\kappa}{1+\kappa}}}{M_2 T^{\frac{\kappa}{1+\kappa}}}$, мы выводим из (6.58) следующее неравенство

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \le (2+1+2)\,\frac{\sigma_q \mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{\kappa}{1+\kappa}}}{T^{\frac{\kappa}{1+\kappa}}}. \quad (6.59)$$

В σ_q τ-зависимый член имеет скорость убывания $T^{\frac{-2\kappa}{1+\kappa}}$, поэтому мы им пренебрегаем. Далее воспользуемся фактом из работы [34](Remark 3) о том, что $D_{\Psi_p}(x^*,x_0) = \tilde{O}(\|x_0 - x^*\|_p^{\frac{1+\kappa}{\kappa}})$, а также введём $R_k = \mathbb{E}\left[\|\overline{x}_k - x^*\|_p^{\frac{1+\kappa}{\kappa}}\right]^{\frac{\kappa}{1+\kappa}}$
.

При r-growth Предположении 4 мы ограничиваем $\mathbb{E}[f(\overline{x}_T)] - f(x^*)$ с двух сторон

$$\frac{\mu_r}{2}\mathbb{E}\left[\|\overline{x}_T - x^*\|_p^r\right] \le \mathbb{E}[f(\overline{x}_T)] - f(x^*) \le \tilde{O}\left(R_0\frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right). \quad (6.60)$$

В силу неравенства Йенсена, которое мы можем применить, поскольку $r \geq \frac{1+\kappa}{\kappa}$, мы перепишем (6.60), чтобы получить в нём R_1

$$\frac{\mu_r}{2}\mathbb{E}\left[\|\overline{x}_T - x^*\|_p^{\frac{1+\kappa}{\kappa}}\right]^{r/\frac{1+\kappa}{\kappa}} \leq \frac{\mu_r}{2}\mathbb{E}\left[\|\overline{x}_T - x^*\|_p^r\right] \leq \tilde{O}\left(R_0\frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right). \tag{6.61}$$

Выясним, через сколько итераций значение R_0 уменьшится вдвое

$$\frac{\mu_r}{2}R_1^r \leq \tilde{O}\left(R_0\frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right) \leq \frac{\mu_r}{2}\left(\frac{R_0}{2}\right)^r. \tag{6.62}$$

Из правого неравенства (6.62) получаем количество итераций за один рестарт

$$T_1 \geq \tilde{O}\left(\left(\frac{2^{(1+r)}\sigma_q}{\mu_r}\right)^{\frac{1+\kappa}{\kappa}}\frac{1}{R_0^{\frac{(r-1)(1+\kappa)}{\kappa}}}\right).$$

Для удобства определим $A \overset{\text{def}}{=} \frac{2^{(1+r)}\sigma_q}{\mu_r}$.

После T_1 итераций перезапускаем алгоритм с начальной точки $x_0 = \overline{x}_{T_1}$ и $R_k = R_{k-1}/2 = R_0/2^k$.

После рестартов N общее количество итераций T будет

$$\begin{aligned} T = \sum_{k=1}^{N} T_k &= \tilde{O}\left(\frac{A^{\frac{1+\kappa}{\kappa}}}{R_0^{\frac{(r-1)(1+\kappa)}{\kappa}}}\sum_{k=0}^{N-1}2^{k\left(\frac{(r-1)(1+\kappa)}{\kappa}\right)}\right) \\ &= \tilde{O}\left(\frac{A^{\frac{(1+\kappa)}{\kappa}}}{R_0^{\frac{(r-1)(1+\kappa)}{\kappa}}}\left[2^{N\left(\frac{(r-1)(1+\kappa)}{\kappa}\right)} - 1\right]\right). \end{aligned} \tag{6.63}$$

На последнем рестарте мы можем получить зависимость точности от количества рестартов N

$$\begin{aligned} \mathbb{E}[f(x_{\text{final}})] - f(x^*) \leq \varepsilon &= \tilde{O}\left(R_{N-1}\frac{\sigma_q}{T_N^{\frac{\kappa}{1+\kappa}}}\right) \\ &\leq \tilde{O}\left(\frac{\mu_r}{2}\left(\frac{R_{N-1}}{2}\right)^r\right) \leq \tilde{O}\left(\frac{\mu_r}{2}\frac{R_0^r}{2^{(N-1)r}}\right) \end{aligned} \tag{6.64}$$

Следовательно, для получения ε точности необходимо N рестартов и общее количество итераций T, где

$$N = \tilde{O}\left(\frac{1}{r}\log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)\right), \tag{6.65}$$

$$T = \tilde{O}\left(\left[\frac{2^{\frac{r^2+1}{r}}\sigma_q}{\mu_r^{1/r}} \cdot \frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right), \quad T_k = \tilde{O}\left(\left[\frac{\sigma_q 2^{(1+r)}}{\mu_r R_0^{r-1}} 2^{k(r-1)}\right]^{\frac{1+\kappa}{\kappa}}\right). \tag{6.66}$$

При каждом рестарте мы получаем разные границы для абсолютного значения шума. Из формулы T_k из (6.63) получаем ограничение

$$\Delta_k = \tilde{O}\left(\frac{\mu_r^2 R_0^{(2r-1)}}{M_2\sqrt{d}}\frac{1}{2^{k(2r-1)}}\right). \tag{6.67}$$

Следовательно, Δ_k будет наименьшим на последней итерации, когда $k = N$, т.е.

$$\Delta_N = \tilde{O}\left(\frac{\mu_r^{1/r}}{M_2\sqrt{d}}\varepsilon^{(2-1/r)}\right).$$

Шаг 2: Сходимость по мат. ожиданию Алгоритма с Клиппингом

Теперь x_0 в Алгоритме 2 можно выбирать случайным образом.

Аналогично доказательству Теоремы 4.0.2, но с $c^* = \frac{\mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{1}{2}}}{\nu^*}, \nu^* = \mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{1}{2}}\left(\frac{1}{4T\sigma_q^{1+\kappa}}\right)^{\frac{1}{1+\kappa}}$ можно получить из (6.44) следующую оценку

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \Delta\frac{\sqrt{d}}{\tau}\mathcal{D}_\Psi + 2\frac{\sigma_q\mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{1}{2}}}{T^{\frac{\kappa}{1+\kappa}}}. \tag{6.68}$$

При обязательном условии $\Delta \le \frac{\sigma_q^2\mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{1}{2}}}{M_2\sqrt{d}T^{\frac{2\kappa}{1+\kappa}}}$ выбрав $\tau = \frac{\sigma_q\mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{1}{2}}}{M_2T^{\frac{\kappa}{1+\kappa}}}$, мы выводим из (6.68) следующее неравенство

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq (2+1+2)\,\frac{\sigma_q \mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{1}{2}}}{T^{\frac{\kappa}{1+\kappa}}}. \tag{6.69}$$

В σ_q τ-зависимый член имеет скорость убывания $T^{\frac{-2\kappa}{1+\kappa}}$, поэтому мы им пренебрегаем. Далее воспользуемся фактом из работы [34](Remark 3) о том, что $D_{\Psi_p}(x^*,x_0) = \tilde{O}(\|x_0 - x^*\|_p^2)$, а также обозначим $R_k = \mathbb{E}\left[\|\overline{x}_k - x^*\|_p^2\right]^{\frac{1}{2}}$.

При r-growth Предположении 4 мы ограничиваем $\mathbb{E}[f(\overline{x}_T)] - f(x^*)$ с двух сторон

$$\frac{\mu_r}{2}\mathbb{E}\left[\|\overline{x}_T - x^*\|_{q^*}^r\right] \leq \mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq \tilde{O}\left(R_0 \frac{\sigma_q}{T^{\frac{\kappa}{(1+\kappa)}}}\right).$$

В силу неравенства Йенсена, которое мы можем применить, поскольку $r \geq 2$, мы получим

$$\frac{\mu_r}{2}\mathbb{E}\left[\|\overline{x}_T - x^*\|_{q^*}^2\right]^{r/2} \leq \frac{\mu_r}{2}\mathbb{E}\left[\|\overline{x}_T - x^*\|_{q^*}^r\right] \leq \tilde{O}\left(R_0 \frac{\sigma_q}{T^{\frac{\kappa}{(1+\kappa)}}}\right).$$

Следующая часть доказательства такая же, как и в **Шаг** 1, начиная с(6.61). Аналогично, мы получаем те же T_2,N_2 и границы шума из (6.66), (6.65) и (6.67) соответственно.

Шаг 3: Сходимость с высокой вероятностью Алгоритма с Клиппингом

Теперь x_0 в Алгоритме 2 можно выбирать случайным образом.

Важным моментом сходимости по высокой вероятности является контроль конечной вероятности. Пусть количество рестартов равно N_3, если вероятность после каждого рестарта находится в пределах не менее $1 - \delta/N_3$, то конечная вероятность попадания в пределы будет больше, чем $1 - \delta$, что является вероятностью «всех рестартов быть в границах». Обычно $N_3 \sim \log(\frac{1}{\varepsilon})$, поэтому

$$\log\frac{N_3}{1} = \log\log\frac{1}{\varepsilon} \ll \log\frac{1}{\delta}\frac{1}{\varepsilon^{\frac{1+\kappa}{\kappa}}}.$$

Это означает, что мы можем использовать $\log\frac{1}{\delta}$ вместо $\log\frac{N_3}{\delta}$.

Аналогично доказательству Теоремы 4.0.3, но с $c^* = \frac{\mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{1}{2}}}{\nu^*}$, $\nu^* = \left[D_{\Psi_p}(x^*,x_0)\right]^{1/2}\left(\frac{1}{T\sigma_q^{1+\kappa}}\right)^{\frac{1}{1+\kappa}}$ можно получить из (6.56) следующую оценку с вероятностью не менее $1-\delta/N_3$

$$
\begin{aligned}
f(\overline{x}_T) - f(x^*) &\leq 2M_2\tau + \Delta\frac{\sqrt{d}}{\tau}\mathcal{D}_\Psi \\
&+ \frac{\left[D_{\Psi_p}(x^*,x_0)\right]^{1/2}\sigma_q}{2T^{\frac{\kappa}{1+\kappa}}}\left[3 + 8\log\frac{1}{\delta} + 12\log\frac{1}{\tilde{\delta}} + 20\log\frac{4}{\delta} \ + 4\sqrt{2\log\frac{1}{\delta}}\right].
\end{aligned}
$$

Введём следующие обозначения для удобства

$$
\tilde{\delta}^{-1} = \frac{4}{\delta}\left[\log\left(\sqrt{T}\right)+2\right]^2, \beta = \left[3 + 8\log\frac{1}{\delta} + 12\log\frac{1}{\tilde{\delta}} + 20\log\frac{4}{\delta} \ + 4\sqrt{2\log\frac{1}{\delta}}\right.
$$

При обязательном условии $\Delta \leq \frac{\beta^2\sigma_q^2 D^{\frac{1}{2}}_{\Psi_p}(x^*,x_0)}{M_2\sqrt{d}T^{\frac{2\kappa}{1+\kappa}}}$ выбрав $\tau = \frac{\beta\sigma_q D^{\frac{1}{2}}_{\Psi_p}(x^*,x_0)}{M_2 T^{\frac{\kappa}{1+\kappa}}}$, мы получим оценку

$$
f(\overline{x}_T) - f(x^*) \leq (2+1+1)\,\frac{\sigma_q\beta\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{1}{2}}}{T^{\frac{\kappa}{1+\kappa}}}.
$$

В σ_q τ-зависимый член имеет скорость убывания $T^{\frac{-2\kappa}{1+\kappa}}$, поэтому мы им пренебрегаем. Далее воспользуемся фактом из работы [34](Remark 3) о том, что $D_{\Psi_p}(x^*,x_0) = \tilde{O}(\|x_0 - x^*\|_p^2)$, а также обозначим $R_k = \|\overline{x}_k - x^*\|_p$.

При r-growth Предположении 4 мы получаем

$$
\frac{\mu_r}{2}\|\overline{x}_T - x^*\|_p^r \leq f(\overline{x}_T) - f(x^*) \leq \tilde{O}\left(R_0\frac{\sigma_q\beta}{T^{\frac{\kappa}{(1+\kappa)}}}\right).
$$

Следующая часть доказательства такая же, как и в **Шаг** 1, начиная с (6.61) и

$$
A \overset{\text{def}}{=} \frac{2^{(1+r)}\beta\sigma_q}{\mu_r}.
$$

Аналогично, мы получаем T_3,N_3 и границы шума из (6.66), (6.65) и (6.67) соответственно.

$$N = \tilde{O}\left(\frac{1}{r}\log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)\right), \tag{6.70}$$

$$T = \tilde{O}\left(\left[\frac{2^{\frac{r^2+1}{r}}\sigma_q\beta}{\mu_r^{1/r}}\cdot\frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right), \quad T_k = \tilde{O}\left(\left[\frac{\sigma_q\beta 2^{(1+r)}}{\mu_r R_0^{r-1}}2^{k(r-1)}\right]^{\frac{1+\kappa}{\kappa}}\right). \tag{6.71}$$

При каждом рестарте мы получаем разные границы для абсолютного значения шума. Из формулы T_k из (6.71) получаем следующие ограничения

$$\Delta_k = \tilde{O}\left(\frac{\mu_r^2 R_0^{(2r-1)}}{M_2\sqrt{d}}\frac{1}{2^{k(2r-1)}}\right). \tag{6.72}$$

Следовательно, Δ_k будет наименьшим на последней итерации, когда $k = N$, т.е.

$$\Delta_N = \tilde{O}\left(\frac{\mu_r^{1/r}}{M_2\sqrt{d}}\varepsilon^{(2-1/r)}\right).$$

□