Математические заметки

Том 107 выпуск 8 июнь 2020

УДК 519.85

Ускоренный и неускореный стохастический градиентный спуск в модельной общности

Д.М. Двинских, А.И. Тюрин, А.В. Гасников, С.С. Омельченко

В статье описывается новый способ получения оценок скорости сходимости оптимальных методов решения задач гладкой (сильно) выпуклой стохастической оптимизации. Способ базируется на получение результатов стохастической оптимизации на основе результатов о сходимости оптимальных методов в условиях неточных градиентов с малыми шумами неслучайной природы. В отличие от известных ранее результатов в данной работе все оценки получаются в модельной общности.

Библиография: 12 названий.

Ключевые слова: стохастическая оптимизация, ускоренный градиентный спуск, модельная общность, композитная оптимизация

1. Введение

В данной работе рассматривается задача стохастической оптимизации [1, 2, 3]

$$f(x) = \mathbb{E}[f(x,\xi)] \to \min_{x\in Q\subseteq\mathbb{R}^n}, \tag{1.1}$$

где множество Q предполагается выпуклым и замкнутым, ξ — случайная величина, математическое ожидание $\mathbb{E}[f(x,\xi)]$ определено и конечно для любого $x \in Q$, функция $f(x)$ – μ-сильно выпуклая в 2-норме ($\mu \geqslant 0$) и имеющая L-Липшицев градиент, т.е. для всех $x, y \in Q$

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2 \leqslant f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2.$$

Предположим, что есть доступ к $\nabla f(x,\xi)$ – стохастическому градиенту $f(x)$, удовлетворяющему следующим условиям[1] (несмещенность и субгауссовость хвостов

Работа А.И. Тюрина в п. 1 поддержана грантом РФФИ 19-31-90062 Аспиранты, а в п. 3 грантом РФФИ 18-31-20005 мол_а_вед, работа А.В. Гасникова в п. 2 выполнена в рамках Программы фундаментальных исследований НИУ ВШЭ и финансировалось в рамках господдержки ведущих университетов Российской Федерации "5-100". Работа Д.М. Двинских была выполнена при поддержке Министерства науки и высшего образования Российской Федерации (госзадание) № 075-00337-20-03.

распределения, с субгауссовской дисперсией σ^2)

$$\mathbb{E}\left[\nabla f(x,\xi)\right] \equiv \nabla f(x), \mathbb{E}\left[\exp\left(\frac{\|\nabla f(x,\xi) - \mathbb{E}[\nabla f(x,\xi)]\|_2^2}{\sigma^2}\right)\right] \leqslant \exp(1), \tag{1.2}$$

для всех $x \in Q$.

Тогда после N вычислений $\nabla f(x,\xi)$ с большой вероятностью имеем[2] [1, 2, 3]

$$f(x_N) - f(x_*) = \tilde{O}\left(\min\left\{\frac{LR^2}{N^p} + \frac{\sigma R}{\sqrt{N}}, LR^2\exp\left(-\left(\frac{\mu}{L}\right)^{\frac{1}{p}}\frac{N}{2}\right) + \frac{\sigma^2}{\mu N}\right\}\right), \tag{1.3}$$

где x_* – решение задачи (1.1), $R = \|x_0 - x_*\|_2$, x_0 – точка старта, $p = 1$ отвечает стохастическому градиентному спуску, а $p = 2$ ускоренному стохастическому спуску.

С другой стороны известно (см. [1, 2, 4, 5, 6]), что если для задачи (1.1) доступен неточный градиент $\nabla_\delta f(x)$, удовлетворяющий для всех $x, y \in Q$ ослабленному условию L-Липшицевости градиента

$$f(x) + \langle \nabla_\delta f(x), y - x\rangle + \frac{\mu}{2}\|y-x\|_2^2 - \delta_1 \leqslant f(y) \leqslant f(x) + \langle \nabla_\delta f(x), y - x\rangle + \frac{L}{2}\|y-x\|_2^2 + \delta_2, \tag{1.4}$$

то после N вычислений $\nabla_\delta f(x)$ для соответствующих модификаций градиентного и ускоренного градиентного спуска можно получить оценку, аналогичную (1.3)[3]

$$\tilde{O}\left(\min\left\{\frac{LR^2}{N^p} + \delta_1 + N^{p-1}\delta_2, LR^2\exp\left(-\left(\frac{\mu}{L}\right)^{\frac{1}{p}}\frac{N}{2}\right) + \delta_1 + \left(\frac{L}{\mu}\right)^{\frac{p-1}{2}}\delta_2\right\}\right). \tag{1.5}$$

В данной статье подмечается, что результат (1.3) может быть получен[4] из результата (1.5). Более того, сделанное наблюдение, оказывается возможным провести и в модельной общности.

2. Основные результаты

Ограничимся для компактности изложения пояснением перехода от (1.5) к (1.3) для случая $\mu = 0$, и с теми же целями переопределим $R = \max_{x,y\in Q}\|x - y\|_2$ (в действительности, все приведенные далее в этом разделе результаты верны для $R = \|x_0 - x_*\|_2$; показывается аналогично [5]). Будем далее дополнительно предполагать, что в ослабленном условии L-Липшицевости градиента для неточного градиента $\nabla_\delta f(x)$ ошибка δ_1 зависит от y и x:

$$f(x) + \langle \nabla_\delta f(x), y - x\rangle - \delta_1(y,x) \leqslant f(y) \leqslant f(x) + \langle \nabla_\delta f(x), y - x\rangle + \frac{L}{2}\|y-x\|_2^2 + \delta_2, \tag{2.1}$$

[1]Заметим, что для задач минимизации функционалов вида суммы условие ограниченности (субгауссовской) дисперсии может не выполняться даже в очень простых (квадратичных) ситуациях. Как следствие, в общем случае приводимые далее результаты не распространяются на задачи минимизации функционалов вида суммы, в которых в качестве стохастического градиента выбирается градиент случайно выбранного слагаемого [7].

[2]Здесь и далее "с большой вероятностью" – означает с вероятностью $\geqslant 1 - \gamma$, а $\tilde{O}(\cdot)$ означает то же самое, что $O(\cdot)$, только числовой множитель зависит от $\ln(N/\gamma)$.

[3]Для $p = 2$ нужно ввести дополнительные ограничения на δ_1, см. ниже.

[4]В сильно выпуклом случае только в смысле сходимости по математическому ожиданию, без оценки вероятностей больших отклонений.

Первое важное наблюдение заключается, в следующем (доказательство более общего утверждения вынесено в Раздел 3).

ПРЕДПОЛОЖЕНИЕ 1. *Пусть даны две последовательности $\delta_1^k(y,x)$ и δ_2^k ($k \geqslant 0$). Будем предполагать, что*

$$\mathbb{E}\left[\delta_1^k(y,x)|\delta_{1,2}^{k-1}, \delta_{1,2}^{k-2}, ...\right] = 0, \text{ (условная несмещенность)}$$

$\delta_1^k(y,x)$ имеет $\left(\hat{\delta}_1\right)^2$-субгауссовскую условную дисперсию, $\sqrt{\delta_2^k}$ имеет $\hat{\delta}_2$-субгауссовский условный второй момент.

ПРЕДПОЛОЖЕНИЕ 2. *Пусть даны две последовательности $\delta_1^k(x,y)$ и δ_2^k ($k \geqslant 0$). Случайная величина $\delta_1^k(x,y)$ имеет $\left(\hat{\delta}_1^k(x,y)\right)^2$-субгауссовский условный момент такой, что*

1. $\hat{\delta}_1^k(x,y) = \hat{\delta}_1^k(x-y)$ *для всех* $x, y \in Q$.
2. $\hat{\delta}_1^k(\alpha z) \leqslant \alpha\hat{\delta}_1^k(z)$ *для всех* $\alpha \geqslant 0$ *и* $z \in B(0,R)$.
3. $\hat{\delta}_1 < +\infty$, *где* $\hat{\delta}_1 \geqslant \sup_{z\in B(0,R)} \hat{\delta}_1^k(z)$.

ТЕОРЕМА 1. *Пусть для последовательностей $\delta_1^k(y,x)$ и δ_2^k ($k \geqslant 0$) верно Предположение 1, тогда*

1. *После N шагов соответствующей модификации градиентного спуска будет верно следующее неравенство:*

$$\mathbb{E}[f(x_N)] - f(x_*) \leqslant O\left(\frac{LR^2}{N} + \hat{\delta}_2\right).$$

И с большой вероятностью

$$f(x_N) - f(x_*) = \tilde{O}\left(\frac{LR^2}{N} + \frac{\hat{\delta}_1}{\sqrt{N}} + \hat{\delta}_2\right).$$

2. *После N шагов соответствующей модификации ускоренного градиентного спуска будет верно следующее неравенство:*

$$\mathbb{E}[f(x_N)] - f(x_*) \leqslant O\left(\frac{LR^2}{N^2} + N\hat{\delta}_2\right).$$

Если дополнительно верно Предположение 2, то с большой вероятностью

$$f(x_N) - f(x_*) = \tilde{O}\left(\frac{LR^2}{N^2} + \frac{\hat{\delta}_1}{\sqrt{N}} + N\hat{\delta}_2\right).$$

Данная теорема является следствием Теоремы 4 и 6 из Раздела 3 для модели вида $\psi_\delta(y,x) = \langle \nabla_\delta f(x), y - x\rangle$.

Если в качестве $\nabla_\delta f(x)$ взять $\nabla f(x,\xi)$, тогда будет выполнено неравенство (2.1) с $\delta_1(y,x) = \langle \nabla f(x) - \nabla f(x,\xi), y - x\rangle$, $\delta_2 = \frac{1}{2L}\|\nabla f(x,\xi) - \nabla f(x)\|_2^2$, с $L := 2L$. Чтобы это понять, достаточно заметить (первое важное наблюдение), что

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

Более того, для $\delta_1(y,x)$ и δ_2 верно (см. обозначения Предположения 1), что $\hat{\delta}_1 = O(\sigma R)$ и $\hat{\delta}_2 = O(\sigma^2/L)$, и для $\delta_1(y,x)$ верно Предположение 2.

Сделанное наблюдение позволяет с помощью Теоремы 1 получить, например, что с большой вероятностью

$$f(x_N) - f(x_*) = \tilde{O}\left(\frac{LR^2}{N^p} + \frac{\sigma R}{\sqrt{N}} + N^{p-1}\frac{\sigma^2}{L}\right), \tag{2.2}$$

причем

$$\mathbb{E}[f(x_N)] - f(x_*) = O\left(\frac{LR^2}{N^p} + N^{p-1}\frac{\sigma^2}{L}\right).$$

Отметим также возможность при $p = 1$ выбора шага h в базовом детерминированном градиентном спуске меньше чем $1/L$. В этом случае оценка будет иметь вид

$$\mathbb{E}[f(x_N)] - f(x_*) = O\left(\frac{R^2}{hN} + h\sigma^2\right).$$

Минимизируя правую часть по h, получим $h = R/\left(\sigma\sqrt{N}\right)$ и

$$\mathbb{E}[f(x_N)] - f(x_*) = O\left(\frac{\sigma R}{\sqrt{N}}\right).$$

Аналогичные оценки можно выписать и в категориях больших отклонений.

Вторым важным наблюдением является следующая теорема (см., например, [5]).

Теорема 2. *(Батчинг) Пусть $\{\xi^l\}_{l=1}^r$ – независимые одинаково распределенные случайные величины (также как случайная величина ξ, которая имеет субгауссовскую дисперсию σ^2). Тогда для σ_r^2 – субгауссовской дисперсии*

$$\nabla^r f\left(x, \{\xi\}_{l=1}^r\right) = \frac{1}{r}\sum_{l=1}^{r} \nabla f(x, \xi^l),$$

справедлива оценка $\sigma_r^2 = O\left(\sigma^2/r\right)$.

Для обоснования перехода от (1.5) к (1.3) положим в (2.1)

$$\nabla_\delta f(x) = \nabla^r f\left(x, \{\xi\}_{l=1}^r\right)$$

и подберем должным образом r. Для подбора r потребуем, чтобы правая часть в оценке (1.3) была равна ε (желаемой точности решения задачи по функции). Чтобы добиться этого исходя из формулы (2.2) согласно теореме 2 нужно выбрать r так, чтобы все слагаемые в (2.2) были порядка ε. То есть

$$\left(\frac{LR^2}{N^p}\right) \simeq \varepsilon,\ \frac{\sigma R}{\sqrt{N}} \simeq \varepsilon,\ N^{p-1}\frac{\sigma^2}{L} \simeq \varepsilon.$$

Получается переопределенная система уравнений на N, r, которая, тем не менее, оказывается совместной $r = \tilde{O}\left(\frac{\sigma^2}{L\varepsilon}\left(\frac{LR^2}{\varepsilon}\right)^{\frac{p-1}{p}}\right)$. При этом,

$$\text{число итераций алгоритма – } N = \tilde{O}\left(\left(\frac{LR^2}{\varepsilon}\right)^{\frac{1}{p}}\right),$$

а число вычислений $\nabla f(x,\xi)$ – $\tilde{O}\left(\max\left\{\left(\frac{LR^2}{\varepsilon}\right)^{\frac{1}{p}}, \frac{\sigma^2R^2}{\varepsilon^2}\right\}\right) = \tilde{O}\left(\left(\frac{LR^2}{\varepsilon}\right)^{\frac{1}{p}} + \frac{\sigma^2R^2}{\varepsilon^2}\right)$. Данные оценки в точности соответствуют тому, что можно получить с помощью батчинга из оценки (1.3). Отметим, что при $p = 2$ данные оценки оптимальны как по числу итераций, так и по числу параллельно вычисляемых стохастических градиентов на каждой итерации [8].

3. Модельная общность

Результаты раздела 2 можно воспроизвести и в модельной общности [1, 4, 6]. Будем говорить, что функция $\psi_\delta(y,x)$ является (δ,L)-моделью целевой функции $f(x)$, если для всех $x,y\in Q$ функция $\psi_\delta(y,x)$ – выпукла по y, $\psi_\delta(x,x)\equiv 0$,

$$f(x)+\psi_\delta(y,x)+\frac{\mu}{2}\|y-x\|_2^2-\delta_1\leqslant f(y)\leqslant f(x)+\psi_\delta(y,x)+\frac{L}{2}\|y-x\|_2^2+\delta_2. \quad (3.1)$$

Для задач композитной оптимизации (см., например, [1, 11]), в которых целевая функция имеет вид $F(x)=f(x)+h(x)$, где $h(x)$ достаточна простая функция, для которой доступен субградиент, а функция $f(x)$ имеет L-Липшицев градиент, и для нее доступен только стохастический градиент $\nabla f(x,\xi)$, в качестве модели можно взять $\psi_\delta(y,x)=\langle\nabla^r f\left(x,\{\xi\}_{l=1}^r\right),y-x\rangle+h(y)-h(x)$. Тогда аналогично разделу 2, получим, что в (3.1) можно положить $L:=2L$, $\hat{\delta}_1=O(\sigma R/r)$, $\hat{\delta}_2=O(\sigma^2/(Lr))$. Это наблюдение позволяет перенести все результаты раздела 2 на задачи стохастической композитной оптимизации.

Введем следующее предположение.

Предположение 3. *Пусть даны две последовательности δ_1^k и δ_2^k $(k\geqslant 0)$. Будем предполагать, что имеется некоторая константа $\tilde{\delta}_1$ такая, что*

$$\mathbb{E}\left[\delta_1^k|\delta_{1,2}^{k-1},\delta_{1,2}^{k-2},...\right]\leqslant\tilde{\delta}_1,$$

δ_1^k имеет $\left(\hat{\delta}_1\right)^2$-субгауссовский условный второй момент, $\sqrt{\delta_2^k}$ имеет $\hat{\delta}_2$-субгауссовский условный второй момент.

Отметим, что Предположение 3 является более общим, чем Предположение 1.

Обозначим $q=1-\frac{\mu}{L}$. Представим градиентный и быстрый градинетный метод в модельной общности (Алгоритм 1 и 2). В Разделах 3.1 и 3.2 представлены теоремы сходимости и соотвествующие доказательства.

3.1. Градиентный метод с моделью.

Лемма 1. *Пусть $\psi(x)$ — выпуклая функция и*

$$y=\arg\min_{x\in Q}\{\psi(x)+\tfrac{\beta}{2}\|x-z\|_2^2+\tfrac{\gamma}{2}\|x-u\|_2^2\},$$

где $\beta\geqslant 0$ и $\gamma\geqslant 0$. Тогда

$$\psi(x)+\tfrac{\beta}{2}\|x-z\|_2^2+\tfrac{\gamma}{2}\|x-u\|_2^2 \geqslant \psi(y)+\tfrac{\beta}{2}\|y-z\|_2^2+\tfrac{\gamma}{2}\|y-u\|_2^2+\tfrac{\beta+\gamma}{2}\|x-y\|_2^2,\ \ \forall x\in Q.$$

Доказательство. *Из критерия оптимальности следует, что:*

$$\exists g\in\partial\psi(y),\ \ \langle g+\tfrac{\beta}{2}\nabla_y\|y-z\|_2^2+\tfrac{\gamma}{2}\nabla_y\|y-u\|_2^2,x-y\rangle\geqslant 0,\ \ \forall x\in Q.$$

Algorithm 1 Градиентный метод

Input: Начальная точка x_0, константа сильной выпуклости $\mu \geqslant 0$, константа липшевости градиента $L > 0$.

for $k \geqslant 0$ **do**

$$\phi_{k+1}(x) := \psi_{\delta_k}(x, x_k) + \frac{L}{2}\|x - x_k\|_2^2,$$
$$x_{k+1} := \arg\min_{x \in Q} \phi_{k+1}(x). \tag{3.2}$$

end for

Output: $y_N = \frac{1}{\sum_{i=1}^N q^{N-i}} \sum_{i=1}^N q^{N-i} x_i$ и

Algorithm 2 Быстрый градиентный метод

Input: Начальная точка x_0, константа сильноый выпуклости $\mu \geqslant 0$, константа липшевости градиента $L > 0$.

Set $y_0 := x_0$, $u_0 := x_0$, $\alpha_0 := 0$, $A_0 := \alpha_0$

for $k \geqslant 0$ **do**

Константа α_{k+1} — это наибольший корень уравнения

$$A_{k+1}(1 + A_k\mu) = L\alpha_{k+1}^2, \quad A_{k+1} := A_k + \alpha_{k+1}. \tag{3.3}$$

$$y_{k+1} := \frac{\alpha_{k+1}u_k + A_k x_k}{A_{k+1}}. \tag{3.4}$$

$$\phi_{k+1}(x) = \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1}) + \frac{1 + A_k\mu}{2}\|x - u_k\|_2^2 + \frac{\alpha_{k+1}\mu}{2}\|x - y_{k+1}\|_2^2,$$
$$u_{k+1} := \arg\min_{x \in Q} \phi_{k+1}(x). \tag{3.5}$$

$$x_{k+1} := \frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}}. \tag{3.6}$$

end for

Output: x_N,

Из $\beta+\gamma$–сильной выпуклости $\psi(x)+\frac{\beta}{2}\|x-z\|_2^2+\frac{\gamma}{2}\|x-u\|_2^2$ получаем, что

$$\psi(x)+\tfrac{\beta}{2}\|x-z\|_2^2+\tfrac{\gamma}{2}\|x-u\|_2^2 \geqslant \psi(y)+\tfrac{\beta}{2}\|y-z\|_2^2+\tfrac{\gamma}{2}\|y-u\|_2^2 + \langle g+\tfrac{\beta}{2}\nabla_y\|y-u\|_2^2+\tfrac{\gamma}{2}\nabla_y\|y-z\|_2^2, x-y\rangle+\tfrac{\beta+\gamma}{2}\|x-y\|_2^2$$

Последние два неравенства доказывают лемму.

Теорема 3. *После N шагов Алгоритма 1 будет верно следующее неравенство:*

$$f(y_N)-f(x_*)\leqslant \min\left\{\frac{LR^2}{2N}, \frac{LR^2}{2}\exp\left(-\frac{\mu}{L}N\right)\right\}+\frac{1}{\sum_{i=1}^N q^{N-i}}\sum_{i=1}^N q^{N-i}(\delta_1^{i-1}+\delta_2^{i-1}).$$

Доказательство. *Из Определения 3.1 получаем:*

$$f(x_N)\leqslant f(x_{N-1})+\psi_{\delta_{N-1}}(x_N,x_{N-1})+\frac{L}{2}\|x_N-x_{N-1}\|_2^2+\delta_2^{N-1}.$$

Воспользуемся Леммой 1 для (3.2)*:*

$$f(x_N)\leqslant f(x_{N-1})+\psi_{\delta_{N-1}}(x,x_{N-1})+\frac{L}{2}\|x-x_{N-1}\|_2^2-\frac{L}{2}\|x-x_N\|_2^2+\delta_2^{N-1}.$$

Воспользуемся левым неравенством из (3.1)*:*

$$f(x_N)\leqslant f(x)+\frac{L-\mu}{2}\|x-x_{N-1}\|_2^2-\frac{L}{2}\|x-x_N\|_2^2+\delta_1^{N-1}+\delta_2^{N-1}. \tag{3.7}$$

Перепишем неравенство для $x=x_*$:

$$\frac{1}{2}\|x_*-x_N\|_2^2\leqslant\frac{1}{L}\left(f(x_*)-f(x_N)+\delta_1^{N-1}+\delta_2^{N-1}\right)+\frac{q}{2}\|x_*-x_{N-1}\|_2^2.$$

Рекусрсивно получаем, что

$$\frac{1}{2}\|x_*-x_N\|_2^2\leqslant\sum_{i=1}^N\left(\frac{q^{N-i}}{L}(f(x_*)-f(x_i)+\delta_1^{i-1}+\delta_2^{i-1})\right)+\frac{q^N}{2}\|x_*-x_0\|_2^2.$$

Учитывая, что $\frac{1}{2}\|x_*-x_N\|_2^2\geqslant 0$ и определение y_N, мы получим:

$$\begin{aligned}\frac{q^N}{2}\|x_*-x_0\|_2^2 &\geqslant \sum_{i=1}^N\left(\frac{q^{N-i}}{L}(f(x_i)-f(x_*)-\delta_1^{i-1}-\delta_2^{i-1})\right)\\ &\geqslant (f(y_N)-f(x_*))\sum_{i=1}^N\frac{q^{N-i}}{L}-\frac{1}{L}\sum_{i=1}^N q^{N-i}(\delta_1^{i-1}+\delta_2^{i-1}).\end{aligned}$$

Разделим обе части последнего неравенства на $\sum_{i=1}^N\frac{q^{N-i}}{L}$:

$$f(y_N)-f(x_*)\leqslant\frac{\frac{q^N}{2}}{\sum_{i=1}^N\frac{q^{N-i}}{L}}\|x_*-x_0\|_2^2+\frac{1}{\sum_{i=1}^N q^{N-i}}\sum_{i=1}^N q^{N-i}(\delta_1^{i-1}+\delta_2^{i-1}).$$

Используя то, что $\sum_{i=1}^N \frac{q^{N-i}}{L} \geqslant \frac{1}{L}$ *и* $q^{N-i} \geqslant q^N$ *для всех* $i \geqslant 0$*, мы получим неравенство:*

$$f(y_N) - f(x_*) \leqslant \frac{L}{2} \min\left\{q^N, \frac{1}{N}\right\} \|x_* - x_0\|_2^2 + \frac{1}{\sum_{i=1}^N q^{N-i}} \sum_{i=1}^N q^{N-i}(\delta_1^{i-1} + \delta_2^{i-1}).$$

Данное неравенство и $q^N \leqslant \exp(-\frac{\mu}{L}N)$ *завершают доказательство теоремы.*

Рассмотрим следствие Теоремы 3.

Теорема 4. *Пусть для последовательностей* δ_1^k *и* δ_2^k *(*$k \geqslant 0$*) верно Предположение 3, тогда после* N *шагов Алгоритма 1 будет верно следующее неравенство:*

$$\mathbb{E}[f(y_N)] - f(x_*) \leqslant \min\left\{\frac{LR^2}{2N}, \frac{LR^2}{2}\exp\left(-\frac{\mu}{L}N\right)\right\} + \tilde{\delta}_1 + O(\hat{\delta}_2). \tag{3.8}$$

Если предположить, что $\mu = 0$*, то с большой вероятностью*

$$f(y_N) - f(x_*) = \tilde{O}\left(\frac{LR^2}{N} + \frac{\hat{\delta}_1}{\sqrt{N}} + \tilde{\delta}_1 + \hat{\delta}_2\right).$$

Доказательство. *Первое неравенство можно получить используя стандартные неравенства для моментов субгуассовских случайных величин* [9]*. Для второго неравенства надо заметить, что в случае* $\mu = 0$ *выполнено равенство:*

$$\frac{1}{\sum_{i=1}^N q^{N-i}} \sum_{i=1}^N q^{N-i}(\delta_1^{i-1} + \delta_2^{i-1}) = \frac{1}{N}\sum_{i=1}^N (\delta_1^{i-1} + \delta_2^{i-1}).$$

Для последнего слагаемого надо воспользоваться неравенствами концентрации для субгауссовских и субэкспоненциальных случайных величин [9].

3.2. Быстрый градиентный метод с моделью. В случае быстрого градиентного метода нам понадобится изменить определение модели функции. Будем говорить, что функция $\psi_\delta(y, x)$ является (δ, L)-моделью целевой функции $f(x)$, если для всех $x, y \in Q$ функция $\psi_\delta(y, x)$ – выпукла по y, $\psi_\delta(x, x) \equiv 0$,

$$f(x) + \psi_\delta(y, x) + \frac{\mu}{2}\|y - x\|_2^2 - \delta_1(y, x) \leqslant f(y) \leqslant f(x) + \psi_\delta(y, x) + \frac{L}{2}\|y - x\|_2^2 + \delta_2. \tag{3.9}$$

Отметим, что теперь мы в общем случае предполагаем, что δ_1 является функцией от двух аргументов $y, x \in Q$.

Лемма 2. *Для всех* $x \in Q$ *выполнено неравенство*

$$\begin{aligned} &A_{k+1}f(x_{k+1}) - A_k f(x_k) + \frac{1 + A_{k+1}\mu}{2}\|x - u_{k+1}\|_2^2 - \frac{1 + A_k\mu}{2}\|x - u_k\|_2^2 \\ &\leqslant \alpha_{k+1}f(x) + A_k\delta_1^k(x_k, y_{k+1}) + \alpha_{k+1}\delta_1^k(x, y_{k+1}) + A_{k+1}\delta_2^k. \end{aligned}$$

Доказательство. *Воспользуемся определением* (3.9)*:*

$$f(x_{k+1}) \leqslant f(y_{k+1}) + \psi_{\delta_k}(x_{k+1}, y_{k+1}) + \frac{L}{2}\|x_{k+1} - y_{k+1}\|_2^2 + \delta_2^k.$$

Из (3.6) *и* (3.4) *для последовательностей* x_{k+1} *и* y_{k+1} *мы получим, что*

$$\begin{aligned} f(x_{k+1}) &\leqslant f(y_{k+1}) + \psi_{\delta_k}\left(\frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}}, y_{k+1}\right) \\ &\quad + \frac{L}{2}\left\|\frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}} - y_{k+1}\right\|_2^2 + \delta_1^k \\ &= f(y_{k+1}) + \psi_{\delta_k}\left(\frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}}, y_{k+1}\right) + \frac{L\alpha_{k+1}^2}{2A_{k+1}^2}\left\|u_{k+1} - u_k\right\|_2^2 + \delta_2^k. \end{aligned}$$

Так как модель $\psi_{\delta_k}(\cdot, y_{k+1})$ *выпуклая, то*

$$\begin{aligned} f(x_{k+1}) &\leqslant \frac{A_k}{A_{k+1}}\left(f(y_{k+1}) + \psi_{\delta_k}(x_k, y_{k+1})\right) + \frac{\alpha_{k+1}}{A_{k+1}}\left(f(y_{k+1}) + \psi_{\delta_k}(u_{k+1}, y_{k+1})\right) \\ &\quad + \frac{L\alpha_{k+1}^2}{2A_{k+1}^2}\left\|u_{k+1} - u_k\right\|_2^2 + \delta_2^k. \end{aligned}$$

Из (3.3) *для последовательности* α_{k+1} *будет верно:*

$$\begin{aligned} f(x_{k+1}) &\leqslant \frac{A_k}{A_{k+1}}\left(f(y_{k+1}) + \psi_{\delta_k}(x_k, y_{k+1})\right) + \frac{\alpha_{k+1}}{A_{k+1}}\Big(f(y_{k+1}) + \psi_{\delta_k}(u_{k+1}, y_{k+1}) \\ &\quad + \frac{1 + A_k\mu}{2\alpha_{k+1}}\left\|u_{k+1} - u_k\right\|_2^2\Big) + \delta_2^k. \end{aligned} \tag{3.10}$$

Из Леммы 1 для оптимизационной задачи (3.5) *будет следовать, что*

$$\begin{aligned} &\alpha_{k+1}\psi_{\delta_k}(u_{k+1}, y_{k+1}) + \frac{1 + A_k\mu}{2}\left\|u_{k+1} - u_k\right\|_2^2 + \frac{\alpha_{k+1}\mu}{2}\left\|u_{k+1} - y_{k+1}\right\|_2^2 \\ &\quad + \frac{1 + A_{k+1}\mu}{2}\left\|x - u_{k+1}\right\|_2^2 \\ &\leqslant \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1}) + \frac{1 + A_k\mu}{2}\left\|x - u_k\right\|_2^2 + \frac{\alpha_{k+1}\mu}{2}\left\|x - y_{k+1}\right\|_2^2. \end{aligned}$$

Так как $\frac{1}{2}\left\|u_{k+1} - y_{k+1}\right\|_2^2 \geqslant 0$, *то*

$$\begin{aligned} &\alpha_{k+1}\psi_{\delta_k}(u_{k+1}, y_{k+1}) + \frac{1 + A_k\mu}{2}\left\|u_{k+1} - u_k\right\|_2^2 \\ &\leqslant \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1}) + \frac{1 + A_k\mu}{2}\left\|x - u_k\right\|_2^2 \\ &\quad - \frac{1 + A_{k+1}\mu}{2}\left\|x - u_{k+1}\right\|_2^2 + \frac{\alpha_{k+1}\mu}{2}\left\|x - y_{k+1}\right\|_2^2. \end{aligned} \tag{3.11}$$

Объединим неравенства (3.10) *и* (3.11), *тогда*

$$\begin{aligned} f(x_{k+1}) &\leqslant \frac{A_k}{A_{k+1}}\left(f(y_{k+1}) + \psi_{\delta_k}(x_k, y_{k+1})\right) \\ &\quad + \frac{\alpha_{k+1}}{A_{k+1}}\Big(f(y_{k+1}) + \psi_{\delta_k}(x, y_{k+1}) + \frac{\mu}{2}\left\|x - y_{k+1}\right\|_2^2 \\ &\quad + \frac{1 + A_k\mu}{2\alpha_{k+1}}\left\|x - u_k\right\|_2^2 - \frac{1 + A_{k+1}\mu}{2\alpha_{k+1}}\left\|x - u_{k+1}\right\|_2^2\Big) + \delta_2^k. \end{aligned}$$

Воспользуемся левым неравенством из (3.9)*:*

$$f(x_{k+1}) \leqslant \frac{A_k}{A_{k+1}} f(x_k) + \frac{\alpha_{k+1}}{A_{k+1}} f(x) + \frac{1 + A_k\mu}{2A_{k+1}} \|x - u_k\|_2^2 - \frac{1 + A_{k+1}\mu}{2A_{k+1}} \|x - u_{k+1}\|_2^2 + \frac{A_k}{A_{k+1}} \delta_1^k(x_k, y_{k+1}) + \frac{\alpha_{k+1}}{A_{k+1}} \delta_1^k(x, y_{k+1}) + \delta_2^k.$$

Данное неравенство завершает доказательство леммы.

Теорема 5. *После N шагов Алгоритма 2 будет верно следующее неравенство:*

$$f(x_N) - f(x_*) \leqslant \frac{R^2}{2A_N} + \frac{1}{A_N} \sum_{k=0}^{N-1} A_k \delta_1^k(x_k, y_{k+1}) + \frac{1}{A_N} \sum_{k=0}^{N-1} \alpha_{k+1} \delta_1^k(x_*, y_{k+1}) + \frac{1}{A_N} \sum_{k=0}^{N-1} A_{k+1} \delta_2^k.$$

Доказательство. *Суммируя неравентсва из Леммы 2 для k от 0 и $N-1$ и, взяв $x = x_*$, мы получим, что*

$$A_N f(x_N) \leqslant A_N f(x_*) + \frac{1}{2} \|x_* - u_0\|_2^2 - \frac{1 + A_N\mu}{2} \|x_* - u_N\|_2^2 + \sum_{k=0}^{N-1} A_k \delta_1^k(x_k, y_{k+1}) + \sum_{k=0}^{N-1} \alpha_{k+1} \delta_1^k(x_*, y_{k+1}) + \sum_{k=0}^{N-1} A_{k+1} \delta_2^k.$$

Так как $\frac{1+A_N\mu}{2} \|x_* - u_N\|_2^2 \geqslant 0$, то

$$A_N f(x_N) - A_N f(x_*) \leqslant \frac{1}{2} \|x_* - u_0\|_2^2 + \sum_{k=0}^{N-1} A_k \delta_1^k(x_k, y_{k+1}) + \sum_{k=0}^{N-1} \alpha_{k+1} \delta_1^k(x_*, y_{k+1}) + \sum_{k=0}^{N-1} A_{k+1} \delta_2^k.$$

Последнее неравенство доказывает теорему.

Лемма 3. *Для всех $N \geqslant 1$,*

$$\frac{1}{A_N} \leqslant \min\left\{ \frac{4L}{N^2}, 2L \exp\left(-\frac{N-1}{2} \sqrt{\frac{\mu}{L}} \right) \right\}.$$

Результат леммы можно получить по аналогии с [2], [12] (см. замечание 5.11.).

Далее нам будут полезны следующие предположения. Похожее на предположение 4 условие на детерминированный шум в градиенте можно встретить в работе [10].

Предположение 4. *Пусть даны две последовательности $\delta_1^k(x, y)$ и δ_2^k ($k \geqslant 0$). Случайная величина $\delta_1^k(x, y)$ имеет такое условное математическое ожидание, что*

1. $\mathbb{E}\left[\delta_1^k(x,y)|\delta_1^{k-1}(x,y),\delta_2^{k-1},\delta_1^{k-2}(x,y)\dots\right] \leqslant \tilde{\delta}_1^k(x,y)\ \forall x,y\in Q$, *где* $\tilde{\delta}_1^k(x,y)$ *есть неслучайная функция от* x *и* y.
2. $\tilde{\delta}_1^k(x,y)=\tilde{\delta}_1^k(x-y)$ *для всех* $x,y\in Q$.
3. $\tilde{\delta}_1^k(\alpha z)\leqslant\alpha\tilde{\delta}_1^k(z)$ *для всех* $\alpha\geqslant 0$ *и* $z\in B(0,R)$.
4. $\tilde{\delta}_1<+\infty$, *где* $\tilde{\delta}_1\geqslant\sup_{z\in B(0,R)}\tilde{\delta}_1^k(z)$.

Теорема 6. *Пусть для последовательностей* $\delta_1^k(x,y)$ *и* δ_2^k *(*$k\geqslant 0$*) верно Предположение 3 и 4 для любых* $x,y\in Q$, *тогда после* N *шагов Алгоритма 2 будет верно следующее неравенство:*

$$\mathbb{E}[f(x_N)]-f(x_*)\leqslant\min\left\{\frac{4LR^2}{N^2},2LR^2\exp\left(-\frac{N-1}{2}\sqrt{\frac{\mu}{L}}\right)\right\}+\tilde{\delta}_1 + O\left(\min\left\{N,\sqrt{\frac{L}{\mu}}\right\}\hat{\delta}_2\right).$$

Предположим дополнительно, что $\mu=0$, *и для последовательности* $\delta_1^k(x,y)$ *выполнено Предположение 2, тогда с большой вероятностью*

$$f(x_N)-f(x_*)=\tilde{O}\left(\frac{LR^2}{N^2}+\frac{\hat{\delta}_1}{\sqrt{N}}+\tilde{\delta}_1+N\hat{\delta}_2\right). \tag{3.12}$$

Доказательство. *Первое неравенство получается из тех же соображений, что и в доказательстве Теоремы 4 с учетом того, что*

$$\frac{1}{A_N}\sum_{k=0}^{N-1}A_{k+1}\leqslant O\left(\min\left\{N,\sqrt{\frac{L}{\mu}}\right\}\right)$$

(см. [2], *Лемма 5.11) и серии неравенств:*

$$A_k\tilde{\delta}_1^k(x_k,y_{k+1})=A_k\tilde{\delta}_1^k(x_k-y_{k+1})=A_k\tilde{\delta}_1^k\left(\frac{\alpha_{k+1}}{A_k}(y_{k+1}-u_k)\right)$$
$$=\alpha_{k+1}\tilde{\delta}_1^k(y_{k+1}-u_k)\leqslant\alpha_{k+1}\tilde{\delta}_1.$$

Во втором переходе мы воспользовались (3.4). *В предпоследнем и последнем переходе использовали Предположение 4. Теперь докажем* (3.12). *Для доказательства неравенства*

$$\frac{1}{A_N}\sum_{k=0}^{N-1}A_{k+1}\delta_2^k\leqslant\tilde{O}(N\hat{\delta}_2)$$

нужно воспользоваться неравенством концентрации для субэкспоненциальных случайных величин. Чтобы показать неравенство

$$\frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\delta_1^k(x_*,y_{k+1})\leqslant\tilde{O}\left(\frac{\hat{\delta}_1}{\sqrt{N}}\right)$$

нужно воспользоваться неравенством концентрации для субгауссовских случайных величин. Неравенство

$$\frac{1}{A_N}\sum_{k=0}^{N-1}A_k\delta_1^k(x_k,y_{k+1})\leqslant\tilde{O}\left(\frac{\hat{\delta}_1}{\sqrt{N}}\right)$$

доказывается аналогично, как и предыдущее, но с учетом того, что

$$A_k\hat{\delta}_1^k(x_k, y_{k+1}) = A_k\hat{\delta}_1^k(x_k - y_{k+1}) = A_k\hat{\delta}_1^k\left(\frac{\alpha_{k+1}}{A_k}(y_{k+1} - u_k)\right)$$
$$\leqslant \alpha_{k+1}\hat{\delta}_1^k(y_{k+1} - u_k) \leqslant \alpha_{k+1}\hat{\delta}_1.$$

Во втором равенстве мы воспользовались (3.4).

3.2.1. *Максимум гладких функций.* Рассмотрим следующую задачу:

$$F(x) := \max_{1\leqslant i\leqslant m} f_i(x) \to \min_{x\in Q\subseteq\mathbb{R}^n}.$$

Будем предполагать, что $f_i(x)$ ($i \in [1, m]$) выпуклые и имеют L-Липшицев градиент, т.е. для всех $x, y \in Q$

$$f_i(x) + \langle\nabla f_i(x), y - x\rangle \leqslant f_i(y) \leqslant f_i(x) + \langle\nabla f_i(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2. \tag{3.13}$$

Выберем в качестве модели функции $F(x)$:

$$\psi_\delta(y, x) = \max_{1\leqslant i\leqslant m}\{f_i(x) + \langle\nabla f_i(x, \xi_i), y - x\rangle\} - F(x),$$

где $f_i(x, \xi_i)$ — независимые стохастические градиенты функций f_i, для которых выполнены условия (1.2). Из (3.13) будет выполнено левое неравенство

$$F(y) \geqslant F(x) + \psi_\delta(y, x) + \max_{1\leqslant i\leqslant m}\{f_i(x) + \langle\nabla f_i(x), y - x\rangle\}$$
$$- \max_{1\leqslant i\leqslant m}\{f_i(x) + \langle\nabla f_i(x, \xi_i), y - x\rangle\}$$
$$\geqslant F(x) + \psi_\delta(y, x) - \max_{1\leqslant i\leqslant m}\{\langle\nabla f_i(x, \xi_i) - \nabla f_i(x), y - x\rangle\}$$

и правое неравенство

$$F(y) \leqslant F(x) + \psi_\delta(y, x) + \max_{1\leqslant i\leqslant m}\{f_i(x) + \langle\nabla f_i(x), y - x\rangle\}$$
$$- \max_{1\leqslant i\leqslant m}\{f_i(x) + \langle\nabla f_i(x, \xi_i), y - x\rangle\} + \frac{L}{2}\|y - x\|_2^2$$
$$\leqslant F(x) + \psi_\delta(y, x) + \frac{1}{2L}\max_{1\leqslant i\leqslant m}\|\nabla f_i(x) - \nabla f_i(x, \xi_i)\|_2^2 + L\|y - x\|_2^2.$$

из (3.9) с

$$\delta_2 = \frac{1}{2L}\max_{1\leqslant i\leqslant m}\|\nabla f_i(x) - \nabla f_i(x, \xi_i)\|_2^2,$$

$$\delta_1(y, x) = \max_{1\leqslant i\leqslant m}\{\langle\nabla f_i(x, \xi_i) - \nabla f_i(x), y - x\rangle\}$$

и $L := 2L$. Из условий (1.2) получаем, что $\delta_1(y, x)$ является субгауссовской, а δ_2 — субэкспоненциальной случайной величиной, так как максимум субгауссовских (субэкспоненциальных) случайных величин есть субгауссовская (субэкспоненциальная) случайная величина. Более того, будут выполнены Предположения 2, 3 и 4. Таким образом для текущей задачи с выбранной моделью применимы Теоремы 4 и 6.

СПИСОК ЦИТИРОВАННОЙ ЛИТЕРАТУРЫ

[1] А.В. Гасников, *Современные численные методы оптимизации. Метод универсального градиентного спуска*, МФТИ, 2018.

[2] O. Devolder, *Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization*, PhD Thesis, ICTEAM and CORE, Université Catholique de Louvain, 2013.

[3] G. Lan, *Lectures on optimization. Methods for Machine Learning.* `https://pwp.gatech.edu/guanghui-lan/publications/`.

[4] А.В. Гасников, А.И. Тюрин, "Быстрый градиентный спуск для задач выпуклой минимизации с оракулом, выдающим (δ, L)-модель функции в запрошенной точке", *ЖВМ и МФ*, **59**:7 (2019), 1137–1150.

[5] E. Gorbunov, D. Dvinskikh, A. Gasnikov, *Optimal decentralized distributed algorithms for stochastic convex optimization*, arXiv preprint `arXiv:1911.07363`.

[6] F. Stonyakin et al., *Inexact model: A framework for optimization and variational inequalities*, arXiv preprint `arXiv:1902.00990`.

[7] M. Assran, M. Rabbat, *On the Convergence of Nesterov's Accelerated Gradient Method in Stochastic Settings*, arXiv preprint `arXiv:2002.12414`.

[8] E. Woodwort et al., "Graph oracle models, lower bounds, and gaps for parallel stochastic optimization", *Advances in neural information processing systems*, 2018, 8496–8506.

[9] A. Juditski, A. Nemirovski, *Large deviations of vector-valued martingales in 2-smooth normed spaces*, arXive preprint `arXiv:0809.0813`.

[10] Ф.С. Стонякин, *Адаптивные градиентные методы для некоторых классов задач негладкой оптимизации*, arXiv preprint `arXiv:1911.08425`.

[11] Yu. Nesterov, *Lectures on convex optimization*, Springer, vol. 137., 2018.

[12] Yu. Nesterov, *Gradient methods for minimizing composite objective function*, 2013.

Д.М. Двинских
Weierstrass Institute, Berlin;
Московский физико-технический институт, Москва;
Институт проблем передачи информации РАН
E-mail: `darina.dvinskikh@wias-berlin.de`

А.И. Тюрин
Высшая школа экономики, Москва
E-mail: `alexandertiurin@gmail.com`

А.В. Гасников
Московский физико-технический институт, Москва;
Институт проблем передачи информации РАН;
Высшая школа экономики
E-mail: `gasnikov@yandex.ru`

С.С. Омельченко
Московский физико-технический институт, Москва
E-mail: `sergey.omelchenko@phystech.edu`