Универсальный метод для задач стохастической композитной оптимизации

А.В. Гасников[1,2] gasnikov.av@mipt.ru

Ю.Е. Нестеров[3] yurii.nesterov@uclouvain.be

[1] Институт проблем передачи информации им. А.А. Харкевича Российской академии наук.
127051, Россия, г. Москва, Большой Каретный переулок, д.19 стр. 1
[2] Лаборатория структурных методов анализа данных в предсказательном моделировании (ПреМоЛаб) Факультета управления и прикладной математики Национального исследовательского Университета «Московский физико-технический институт»
141700, Россия, Московская область, г. Долгопрудный, Институтский переулок, д. 9
[3] Center for Operation Research and Econometrics Université Catholique de Louvain.
Voie du Roman Pays 34, L1.03.01 - B-1348 Louvain-la-Neuve (Belgium)

В работе впервые предлагается быстрый градиентный метод для задач гладкой выпуклой оптимизации, требующий всего одну проекцию. Метод имеет наглядную геометрическую интерпретацию, поэтому получил название “метода треугольника” (МТ). В работе также предлагаются: композитный, адаптивный и универсальный вариант МТ. Впервые (на базе МТ) предлагается универсальный метод для сильно выпуклых задач (причем предложенный метод оказался непрерывным по параметру сильной выпуклости гладкой части функционала). Показывается, как универсальный вариант МТ можно применять к задачам стохастической оптимизации.

Метод треугольника для задач композитной оптимизации

Рассматривается задача выпуклой композитной оптимизации

$$F(x) = f(x) + h(x) \to \min_{x \in Q}. \qquad (1)$$

Положим $R^2 = V(x_*, x^0)$, где прокс-расстояние определяется формулой

$$V(x,z) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle;$$

прокс-функция $d(x) \geq 0$ считается сильно выпуклой относительно выбранной нормы $\| \|$, с константой сильной выпуклости ≥ 1; x_* – решение задачи (1) (если решение не единственно, то выбирается то, которое доставляет минимум $V(x_*, x^0)$); $x^0 = y^0 = u^0$ – точка старта итерационного процесса.

Предположение 1. *Для любых* $x, y \in Q$ *имеет место неравенство*

$$\left\|\nabla f\left(y\right)-\nabla f\left(x\right)\right\|_{*} \le L\left\|y-x\right\|.$$

Опишем вариант быстрого градиентного метода для задачи (1) с одной “проекцией”, который мы далее будем называть “метод треугольника” (МТ).

Положим

$$\varphi_0\left(x\right)=V\left(x,y^0\right)+\alpha_0\left[\,f\left(y^0\right)+\left\langle\nabla f\left(y^0\right),x-y^0\right\rangle+h\left(x\right)\right],$$

$$\varphi_{k+1}\left(x\right)=\varphi_k\left(x\right)+\alpha_{k+1}\left[\,f\left(y^{k+1}\right)+\left\langle\nabla f\left(y^{k+1}\right),x-y^{k+1}\right\rangle+h\left(x\right)\right], \qquad (2)$$

$$A_k=\sum_{i=0}^{k}\alpha_i,\ \ \alpha_0=L^{-1},\ A_k=\alpha_k^2 L,\ k=0,1,2,... \qquad (3)$$

Метод Треугольника (Ю.Е. Нестеров, апрель 2016)

$$u^k = \arg\min_{x \in Q} \varphi_k(x),$$

$$y^{k+1} = \frac{\alpha_{k+1} u^k + A_k x^k}{A_{k+1}}, \qquad (4)$$

$$x^{k+1} = \frac{\alpha_{k+1} u^{k+1} + A_k x^k}{A_{k+1}}.$$

Лемма 1. *Последовательность* $\{\alpha_k\}$, *определяемую формулой (3), можно задавать рекуррентно*

$$\alpha_{k+1} = \frac{1}{2L} + \sqrt{\frac{1}{4L^2} + \alpha_k^2}\,.$$

При этом

$$A_k \geq \frac{(k+1)^2}{4L}.$$

Лемма 2. *Пусть справедливо предположение 1. Тогда для любого* $k = 0,1,2,...$ *имеет место неравенство*

$$A_k F\left(x^k\right) \leq \varphi_k^* = \min_{x \in Q} \varphi_k\left(x\right) = \varphi_k\left(u^k\right). \qquad (5)$$

Доказательство. Проведем по индукции. При $k=0$ формула (5) очевидна

$$F\left(x^0\right)=f\left(x^0\right)+h\left(x^0\right)=f\left(y^0\right)+h\left(y^0\right)\le$$

$$\le\underbrace{\frac{\alpha_0}{A_1}}_{=1}\left[f\left(y^0\right)+\underbrace{\left\langle\nabla f\left(y^0\right),u^0-y^0\right\rangle}_{=0}+h\left(y^0\right)\right]+\frac{1}{A_1}\underbrace{V\left(u^0,y^0\right)}_{=0}=\frac{1}{A_1}\varphi_0\left(u^0\right).$$

Итак, пусть формула (5) установлена для k, покажем, что тогда она будет справедлива и для $k+1$. По определению (2)

$$\varphi_{k+1}^*=\min_{x\in Q}\varphi_{k+1}\left(x\right)=\varphi_{k+1}\left(u^{k+1}\right)=$$

$$=\varphi_k\left(u^{k+1}\right)+\alpha_{k+1}\left[f\left(y^{k+1}\right)+\left\langle\nabla f\left(y^{k+1}\right),u^{k+1}-y^{k+1}\right\rangle+h\left(u^{k+1}\right)\right]. \quad (6)$$

Поскольку по предположению индукции $A_k F\left(x^k\right) \le \varphi_k\left(u^k\right)$, и $\varphi_{k+1}(x)$ – сильно выпуклая в $\|\ \|$-норме функция с константой ≥ 1 (это следует из аналогичного свойства функции $V\left(x, y^0\right)$, что, в свою очередь, следует из аналогичного свойства функции $d(x)$), то

$$\varphi_k\left(u^{k+1}\right) \ge \varphi_k\left(u^k\right) + \frac{1}{2}\left\|u^{k+1} - u^k\right\|^2 \ge A_k \cdot \left(f\left(x^k\right) + h\left(x^k\right)\right) + \frac{1}{2}\left\|u^{k+1} - u^k\right\|^2.$$

Из выпуклости $f(x)$ отсюда имеем

$$\varphi_k\left(u^{k+1}\right) \ge A_k f\left(y^{k+1}\right) + \left\langle \nabla f\left(y^{k+1}\right), A_k \cdot \left(x^k - y^{k+1}\right)\right\rangle + A_k h\left(x^k\right) + \frac{1}{2}\left\|u^{k+1} - u^k\right\|^2. (7)$$

Подставляя (7) в (6), получим

$$\varphi_{k+1}^{*} \geq A_{k+1} \cdot \underbrace{\left(\frac{A_k}{A_{k+1}} h\left(x^k\right) + \frac{\alpha_{k+1}}{A_{k+1}} h\left(u^{k+1}\right) \right)}_{\geq h\left(x^{k+1}\right)} + A_{k+1} f\left(y^k\right) +$$

$$+ \left\langle \ \nabla f\left(y^{k+1}\right), \underbrace{\alpha_{k+1} \cdot \left(u^{k+1} - y^{k+1}\right) + A_k \cdot \left(x^k - y^{k+1}\right)}_{= A_{k+1} \cdot \left(x^{k+1} - y^{k+1}\right)} \ \right\rangle + \underbrace{\frac{1}{2} \left\| u^{k+1} - u^k \right\|^2}_{= \frac{A_{k+1}^2}{2\alpha_{k+1}^2} \left\| x^{k+1} - y^{k+1} \right\|^2} \qquad (8)$$

Исходя из выпуклости функции $h\left(x\right)$ и описания МТ (4), формулу (8) можно переписать следующим образом

$$\varphi_{k+1}^{*} \geq A_{k+1} \left[f\left(y^{k+1}\right) + \left\langle \nabla f\left(y^{k+1}\right), x^{k+1} - y^{k+1} \right\rangle + \frac{A_{k+1}}{2\alpha_{k+1}^2} \left\| x^{k+1} - y^{k+1} \right\|^2 + h\left(x^{k+1}\right) \right]. \qquad (9)$$

Из предположения 1 следует, что если $A_{k+1}/\alpha_{k+1}^2 \geq L$, то

$$f\left(y^{k+1}\right)+\left\langle \nabla f\left(y^{k+1}\right), x^{k+1}-y^{k+1}\right\rangle+\frac{A_{k+1}}{2\alpha_{k+1}^2}\left\|x^{k+1}-y^{k+1}\right\|^2 \geq f\left(x^{k+1}\right). \quad (10)$$

Согласно (3) $A_{k+1}/\alpha_{k+1}^2 = L$, поэтому формула (10) имеет место. С помощью формулы (10) формулу (9) можно переписать следующим образом

$$\varphi_{k+1}^{*} \geq A_{k+1}\left[f\left(x^{k+1}\right)+h\left(x^{k+1}\right)\right] = A_{k+1}F\left(x^{k+1}\right).$$

Таким образом, шаг индукции установлен. Следовательно, лемма 2 доказана. ■

Из лемм 1, 2 получаем следующий результат, означающий, что МТ сходится как обычный быстрый градиентный метод (с двумя проекциями), т.е. МТ сходится оптимальным образом для рассматриваемого класса задач.

Теорема 1. *Пусть справедливо предположение 1. Тогда МТ (2) – (4) для задачи (1) сходится согласно оценке*

$$F\left(x^N\right)-\min_{x\in Q}F\left(x\right)\le\frac{4LR^2}{\left(N+1\right)^2}. \qquad (11)$$

Доказательство. Из леммы 2 следует, что (в третьем неравенстве используется выпуклость функции $f(x)$)

$$A_N F(x^N) \le \min_{x \in Q} \left\{ V(x, x^0) + \sum_{k=0}^{N} \alpha_k \left[f(y^k) + \left\langle \nabla f(y^k), x - y^k \right\rangle + h(x) \right] \right\} \le$$

$$\le V(x_*, x^0) + \sum_{k=0}^{N} \alpha_k \underbrace{\left[f(y^k) + \left\langle \nabla f(y^k), x_* - y^k \right\rangle + h(x_*) \right]}_{\le f(x_*) + h(x_*)} \le$$

$$\le V(x_*, x^0) + \sum_{k=0}^{N} \alpha_k F(x_*) = R^2 + A_N F(x_*). \qquad (12)$$

Заметим, что из второго неравенства следует, что если решение задачи (1) x_* не единственно, то можно выбирать то, которое

доставляет минимум $V\left(x_*, x^0\right)$. Именно таким образом возникает R^2 в оценке (12). Для доказательства теоремы осталось подставить нижнюю оценку на A_N из леммы 1 в формулу (12). ■

Замечание 1. В действительности в формуле (12) содержится более сильный результат, чем в формуле (11). А именно, формула (12) еще означает, что МТ – прямо-двойственный метод. Мы не будем здесь подробно на этом останавливаться, отметим лишь, что это свойство позволяет получать эффективные критерии останова для МТ. Критерий останова позволяет не делать предписанного формулой (11) числа итераций и останавливаться раньше (по достижению желаемой точности). Это замечание можно распространить и на все последующее изложение.

Метод треугольника для сильно выпуклых задач композитной оптимизации

Теперь будем считать, что $f(x)$ в задаче (1) обладает следующим свойством.

Предположение 2. *$f(x)$ – μ-сильно выпуклая функция в норме $\|\ \|$, т.е. для любых $x, y \in Q$ имеет место неравенство*

$$f(y)+\langle \nabla f(y), x-y\rangle+\frac{\mu}{2}\|x-y\|^2 \le f(x). \qquad (13)$$

Не ограничивая общности также будем считать, что прокс-расстояние V можно выбрать так, чтобы оно удовлетворяло условию (в евклидовом случае $\omega_n = 1$)

$$\omega_n = \sup_{x,y\in Q} \frac{2V(x,y)}{\|y-x\|^2} = \mathrm{O}(\ln n).$$

Перепишем формулы (2), (3) следующим образом ($\tilde{\mu} = \mu/\omega_n$)

$$\varphi_0(x) = V(x, y^0) + \alpha_0 \left[f(y^0) + \left\langle \nabla f(y^0), x - y^0 \right\rangle + \tilde{\mu} V(x, y^0) + h(x) \right],$$

$$\varphi_{k+1}(x) = \varphi_k(x) + \alpha_{k+1} \left[f(y^{k+1}) + \left\langle \nabla f(y^{k+1}), x - y^{k+1} \right\rangle + \tilde{\mu} V(x, y^k) + h(x) \right], \quad (14)$$

$$A_k = \sum_{i=0}^{k} \alpha_i, \ \alpha_0 = L^{-1}, \ A_{k+1} \cdot (1 + A_k \tilde{\mu}) = \alpha_{k+1}^2 L, \ k = 0, 1, 2, \ldots \quad (15)$$

Сам метод по-прежнему будет иметь вид (4) с

$$x^0 = y^0 = u^0 = \arg\min_{x \in Q} \varphi_0(x).$$

Лемма 3. *Последовательность* $\{\alpha_k\}$, *определяемую формулой (15), можно задавать рекуррентно*

$$\alpha_{k+1} = \frac{1+A_k\tilde{\mu}}{2L} + \sqrt{\frac{1+A_k\tilde{\mu}}{4L^2} + \frac{A_k\cdot(1+A_k\tilde{\mu})}{L}},\ A_{k+1} = A_k + \alpha_{k+1}. \qquad (16)$$

При этом

$$A_k \geq \frac{1}{L}\left(1+\frac{1}{2}\sqrt{\frac{\tilde{\mu}}{L}}\right)^{2k} \geq \exp\left(\frac{k}{2}\sqrt{\frac{\tilde{\mu}}{L}}\right).$$

Лемма 4. *Пусть справедливы предположения 1, 2. Тогда для любого* $k = 0,1,2,\ldots$ *имеет место неравенство*

$$A_k F\left(x^k\right) \leq \varphi_k^* = \min_{x\in Q}\varphi_k\left(x\right) = \varphi_k\left(u^k\right).$$

Доказательство. Доказательство аналогично доказательству леммы 2. В основе лежит неравенство (следующее из ≥ 1-сильной выпуклости $d(x)$ в норме $\|\ \|$)

$$V\left(x, y^k\right) \geq \frac{1}{2}\left\|x - y^k\right\|^2,$$

с помощью которого ключевая формула (8) перепишется следующим образом

$$\varphi_{k+1}^* \geq A_{k+1} \cdot \left(\frac{A_k}{A_{k+1}} h\left(x^k\right) + \frac{\alpha_{k+1}}{A_{k+1}} h\left(u^{k+1}\right) \right) + A_{k+1} f\left(y^k\right) +$$

$$+ \left\langle\ \nabla f\left(y^{k+1}\right), \alpha_{k+1} \cdot \left(u^{k+1} - y^{k+1}\right) + A_k \cdot \left(x^k - y^{k+1}\right)\ \right\rangle + \frac{\left(1 + A_k \tilde{\mu}\right)}{2}\left\|u^{k+1} - u^k\right\|^2.$$

Отличие от формулы (8) в следующем

$$\frac{1}{2}\left\|u^{k+1}-u^{k}\right\|^{2} \to \frac{\left(1+A_k\tilde{\mu}\right)}{2}\left\|u^{k+1}-u^{k}\right\|^{2}.$$

Рассуждая дальше точно также как при доказательстве леммы 2, получим

$$\varphi_{k+1}^{*} \geq A_{k+1}\left[f\left(y^{k+1}\right)+\left\langle \nabla f\left(y^{k+1}\right), x^{k+1}-y^{k+1}\right\rangle + \frac{A_{k+1}\cdot\left(1+A_k\tilde{\mu}\right)}{2\alpha_{k+1}^{2}}\left\|x^{k+1}-y^{k+1}\right\|^{2} + h\left(x^{k+1}\right)\right]. \quad (17)$$

Из предположения 1 следует, что если $A_{k+1}\cdot\left(1+A_k\tilde{\mu}\right)/\alpha_{k+1}^{2} \geq L$, то

$$f\left(y^{k+1}\right)+\left\langle \nabla f\left(y^{k+1}\right), x^{k+1}-y^{k+1}\right\rangle + \frac{A_{k+1}\cdot\left(1+A_k\tilde{\mu}\right)}{2\alpha_{k+1}^{2}}\left\|x^{k+1}-y^{k+1}\right\|^{2} \geq f\left(x^{k+1}\right). \quad (18)$$

Согласно (15) $A_{k+1}\cdot\left(1+A_k\tilde{\mu}\right)\big/\alpha_{k+1}^2=L$, поэтому формула (18) имеет место. С помощью формулы (18) формулу (17) можно переписать следующим образом

$$\varphi_{k+1}^{*}\geq A_{k+1}\left[f\left(x^{k+1}\right)+h\left(x^{k+1}\right)\right]=A_{k+1}F\left(x^{k+1}\right).\ \blacksquare$$

Из лемм 3, 4 получаем следующий результат, означающий, что МТ в сильно выпуклом случае сходится как обычный быстрый градиентный метод (с двумя проекциями), т.е. МТ сходится оптимальным образом для рассматриваемого класса задач.

Теорема 2. *Пусть справедливы предположения 1, 2. Тогда МТ (14), (15), (4) для задачи (1) сходится согласно оценке*

$$F\left(x^N\right)-\min_{x\in Q}F\left(x\right)\le LR^2\exp\left(-\frac{N}{2}\sqrt{\frac{\tilde{\mu}}{L}}\right). \qquad (19)$$

Доказательство. Из леммы 4 следует, что (в третьем неравенстве используется то, что $\tilde{\mu} = \mu / \omega_n$ и сильная выпуклость функции $f(x)$ – см. формулу (13) предположения 2)

$$A_N F(x^N) \le \min_{x \in Q} \left\{ V(x, x^0) + \sum_{k=0}^{N} \alpha_k \left[f(y^k) + \left\langle \nabla f(y^k), x - y^k \right\rangle + \tilde{\mu} V(x, y^k) + h(x) \right] \right\} \le$$

$$\le V(x_*, x^0) + \sum_{k=0}^{N} \alpha_k \underbrace{\left[f(y^k) + \left\langle \nabla f(y^k), x_* - y^k \right\rangle + \frac{\mu}{2} \left\| x_* - y^k \right\|^2 + h(x_*) \right]}_{\le f(x_*) + h(x_*)} \le$$

$$\le V(x_*, x^0) + \sum_{k=0}^{N} \alpha_k F(x_*) = R^2 + A_N F(x_*). \blacksquare \qquad (20)$$

В действительности, выше установлено более сильное утверждение.

Теорема 3. *Пусть справедливы предположения 1, 2. Тогда МТ (14), (16), (4) для задачи (1) сходится согласно оценке*

$$F\left(x^N\right)-\min_{x\in Q}F\left(x\right)\le\min\left\{\frac{4LR^2}{\left(N+1\right)^2},LR^2\exp\left(-\frac{N}{2}\sqrt{\frac{\mu}{L\omega_n}}\right)\right\}. \qquad (21)$$

Теорема 3 означает, что МТ (14), (16), (4) непрерывен по параметру μ. К сожалению, при этом в (16) явно входит этот параметр μ. Если значение этого параметра неизвестно, то с помощью рестартов можно получить оценку (21) увеличив константы не более чем в 4 раза, т.е. число вычислений градиента $\nabla f\left(x\right)$ (обычно именно это является самым затратным в шаге), необходимых для достижения заданной точности, увеличится не более чем в 4 раза. Между сильно выпуклым и просто выпуклым случаями имеется глубокая связь, позволяющая, например, получить оценку (19) с помощью оценки (11) и наоборот. Другими словами, имея эффективные алгоритмы решения выпуклых / сильно выпуклых задач, можно предложить на их базе алгоритмы решения сильно выпуклых / выпуклых задач.

Введем семейство μ-сильно выпуклых в норме $\|\ \|$ задач ($\mu > 0$)

$$F^{\mu}(x) = F(x) + \mu V(x, x^0) \to \min_{x \in Q}. \qquad (22)$$

Теорема 4. *Пусть*

$$\mu \le \frac{\varepsilon}{2V(x_*, y^0)} = \frac{\varepsilon}{2R^2}, \qquad (23)$$

и удалось найти $\varepsilon/2$*-решение задачи (22), т.е. нашелся такой* $x^N \in Q$, *что* $F^{\mu}(x^N) - F_*^{\mu} \le \varepsilon/2$. *Тогда* $F(x^N) - \min\limits_{x \in Q} F(x) = F(x^N) - F_* \le \varepsilon$.

Доказательство. Действительно,

$$F(x^N) - F_* \le F^{\mu}(x^N) - F_* \le F^{\mu}(x^N) - F_*^{\mu} + \varepsilon/2 \le \varepsilon.$$

Здесь использовалось определение F_*^{μ} и формула (23)

$$F_*^{\mu} = \min_{x \in Q}\left\{F(x) + \gamma V(x, x^0)\right\} \le F(x_*) + \gamma V(x_*, x^0) \le F_* + \varepsilon/2. \ \blacksquare$$

Теорема 5. *Пусть функция* $F(x)$ – μ-*сильно выпуклая в норме* $\|\ \|$. *Пусть точка* $x^{\bar{N}}\left(x^0\right)$ *выдается МТ (2) – (4), стартующим с точки* x^0, *после*

$$\bar{N} = \sqrt{\frac{8L\omega_n}{\mu}} \tag{24}$$

итераций. Положим $\left[x^{\bar{N}}\left(x^0\right)\right]^1 = x^{\bar{N}}\left(x^0\right)$ *и определим по индукции*

$$\left[x^{\bar{N}}\left(x^0\right)\right]^{k+1} = x^{\bar{N}}\left(\left[x^{\bar{N}}\left(x^0\right)\right]^k\right),\ k = 1,2,....$$

Тогда

$$F\left(\left[x^{\bar{N}}\right]^k\right) - F_* \le \frac{\mu\left\|x^0 - x_*\right\|^2}{2^{k+1}}. \tag{25}$$

Доказательство. МТ (2) – (4) согласно теореме 1 (см. формулу (11)) после $\bar{N}$ итераций выдает такой $x^{\bar{N}}$, что

$$\frac{\mu}{2}\left\|x^{\bar{N}}-x_*\right\|^2 \le F\left(x^{\bar{N}}\right)-F_* \le \frac{4LV\left(x_*,x^0\right)}{\bar{N}^2}.$$

Отсюда имеем

$$\left\|x^{\bar{N}}-x_*\right\|^2 \le \frac{8LV\left(x_*,y^0\right)}{\mu\bar{N}^2} \le \frac{1}{2}\left\|x^0-x_*\right\|^2\frac{8L\omega_n}{\mu\bar{N}^2}.$$

Поскольку $\bar{N}=\sqrt{\frac{8L}{\mu}\omega_n}$, то $\left\|x^{\bar{N}}-x_*\right\|^2 \le \frac{1}{2}\left\|x^0-x_*\right\|^2$.

Повторяя эти рассуждения, по индукции получим

$$F\left(\left[x^{\bar{N}}\right]^k\right)-F_* \le \left(\frac{1}{2}\right)^k\left\|x^0-x_*\right\|^2\frac{4L\omega_n}{\bar{N}^2}=\frac{\mu\left\|x^0-x_*\right\|^2}{2^{k+1}}. \blacksquare$$

Универсальный метод треугольника

В ряде приложений значение константы L, необходимой МТ для работы (см. формулу (16)), не известно. Однако, как следует из формул (10), (18), знание константы L необязательно, если разрешается на одной итерации запрашивать значение функции в нескольких точках. Опишем соответствующий адаптивный вариант МТ (14), (16), (4) (АМТ).

Положим $A_0^0 = \alpha_0^0 = L_0^0$ – параметр метода и $k = 0,\ j_0 = 0.$

Адаптивный Метод Треугольника

1. $L_{k+1}^0 := L_k^{j_k} / 2,\ j_{k+1} = 0,$

$$u^k = \arg\min_{x \in Q} \varphi_k(x).$$

2. $\alpha_{k+1} := \dfrac{1 + A_k \tilde{\mu}}{2L_{k+1}^{j_{k+1}}} + \sqrt{\dfrac{1 + A_k \tilde{\mu}}{4\left(L_{k+1}^{j_{k+1}}\right)^2} + \dfrac{A_k \cdot (1 + A_k \tilde{\mu})}{L_{k+1}^{j_{k+1}}}},\ A_{k+1} := A_k + \alpha_{k+1},$

$$x^{k+1} := \frac{\alpha_{k+1} u^{k+1} + A_k x^k}{A_{k+1}},\ y^{k+1} := \frac{\alpha_{k+1} u^k + A_k x^k}{A_{k+1}}.$$

До тех пор пока

$$f\left(y^{k+1}\right)+\left\langle \nabla f\left(y^{k+1}\right), x^{k+1}-y^{k+1}\right\rangle+\frac{L_{k+1}^{j_{k+1}}}{2}\left\|x^{k+1}-y^{k+1}\right\|^2 < f\left(x^{k+1}\right),$$

выполнять

$$j_{k+1} := j_{k+1}+1;\ L_{k+1}^{j_{k+1}} := 2^{j_{k+1}} L_k^{j_k}.$$

3. Если не выполнен критерий останова, то $k := k+1$ и **go to** 1.

В качестве критерия останова, например, можно брать условие

$$\left\| x^{k+1} - \arg\min_{x\in Q}\left\{\left\langle \nabla f\left(x^{k+1}\right), x-x^{k+1}\right\rangle+\frac{L_{k+1}^{j_{k+1}}}{2}\left\|x-x^{k+1}\right\|^2\right\}\right\| \le \tilde{\varepsilon}.$$

Теорема 6. *Пусть справедливы предположения 1, 2. Тогда АМТ для задачи (1) сходится согласно оценке*

$$F\left(x^N\right)-\min_{x\in Q}F\left(x\right)\le\min\left\{\frac{8LR^2}{\left(N+1\right)^2},2LR^2\exp\left(-\frac{N}{2}\sqrt{\frac{\mu}{2L\omega_n}}\right)\right\}. \qquad (26)$$

При этом среднее число вычислений значения функции на одной итерации будет ≈ 4*, а градиента функции* ≈ 2.

Доказательство. Нетривиальным в виду оценки (21) и свойства, что все $L_k^{j_k} \le 2L$, представляется только последняя часть формулировки теоремы. Докажем именно её. Оценим общее число обращений за значениями функции (аналогично получается оценка общего числа обращений за значением градиента функции)

$$\sum_{k=1}^{N} 2\left(j_k + 1\right) = \sum_{k=1}^{N} 2\left[\left(j_k - 1\right) + 2\right] = \sum_{k=1}^{N} 2\left[\log_2\left(\frac{L_k^{j_k}}{L_{k-1}^{j_{k-1}}}\right) + 2\right] =$$

$$= 4N + \log_2\left(\frac{L_N^{j_N}}{L_0^0}\right) \le 4N + \log_2\left(\frac{2L}{L_0^0}\right).$$

Деля обе части на N, получим в правой части приблизительно 4. ■

Предположим теперь, что по каким-то причинам невозможно получить точные значения функции и градиента. Тогда соотношение

$$f\left(y^{k+1}\right)+\left\langle \nabla f\left(y^{k+1}\right), x^{k+1}-y^{k+1}\right\rangle+\frac{L_{k+1}^{j_{k+1}}}{2}\left\|x^{k+1}-y^{k+1}\right\|^2 \geq f\left(x^{k+1}\right)$$

может не выполниться не при каком $L_{k+1}^{j_{k+1}}$. Допустим, однако, что при этом имеет место

$$f\left(y^{k+1}\right)+\left\langle \nabla f\left(y^{k+1}\right), x^{k+1}-y^{k+1}\right\rangle+\frac{L}{2}\left\|x^{k+1}-y^{k+1}\right\|^2+\frac{\alpha_{k+1}}{A_{k+1}}\varepsilon \geq f\left(x^{k+1}\right).$$

Тогда заменим в АМТ соответствующую часть шага 2 на

$$f\left(y^{k+1}\right)+\left\langle \nabla f\left(y^{k+1}\right), x^{k+1}-y^{k+1}\right\rangle+\frac{L_{k+1}^{j_{k+1}}}{2}\left\|x^{k+1}-y^{k+1}\right\|^2+\frac{\alpha_{k+1}}{A_{k+1}}\varepsilon \geq f\left(x^{k+1}\right). \quad (27)$$

Теорема 7. *Пусть справедливо предположение 2 и существует такое число* $L > 0$, *что любого* $k = 1, \ldots, N$ *справедливо неравенство*

$$f\left(y^{k+1}\right) + \left\langle \nabla f\left(y^{k+1}\right), x^{k+1} - y^{k+1} \right\rangle + \frac{L}{2}\left\| x^{k+1} - y^{k+1} \right\|^2 + \frac{\alpha_{k+1}}{A_{k+1}}\varepsilon \geq f\left(x^{k+1}\right). \quad (28)$$

Тогда АМТ с (27) для задачи (1) сходится согласно оценке

$$F\left(x^N\right) - \min_{x \in Q} F(x) \leq \min\left\{ \frac{8LR^2}{(N+1)^2}, 2LR^2 \exp\left(-\frac{N}{2}\sqrt{\frac{\mu}{2L\omega_n}} \right) \right\} + \varepsilon. \quad (29)$$

При этом среднее число вычислений значения функции на одной итерации будет ≈ 4, *а градиента функции* ≈ 2

Доказательство. Ключевым элементом в доказательстве является следующее уточнение леммы 4

$$A_k F\left(x^k\right) \le \varphi_k^* + A_k \varepsilon, \qquad (30)$$

из которого будет следовать формула (29). Чтобы доказать (30) будем рассуждать по индукции. База индукции $k = 0$ очевидна. Итак, по предположению индукции

$$A_k F\left(x^k\right) - A_k \varepsilon \le \varphi_k^* = \varphi_k\left(u^k\right),$$

поэтому

$$\varphi_k\left(u^{k+1}\right) \ge \varphi_k\left(u^k\right) + \frac{1 + A_k \tilde{\mu}}{2}\left\|u^{k+1} - u^k\right\|^2 \ge A_k F\left(x^k\right) - A_k \varepsilon + \frac{1 + A_k \tilde{\mu}}{2}\left\|u^{k+1} - u^k\right\|^2.$$

Отсюда

$$\varphi_{k+1}^{*} \geq A_{k+1} \cdot \left(\frac{A_k}{A_{k+1}} h\left(x^k\right) + \frac{\alpha_{k+1}}{A_{k+1}} h\left(u^{k+1}\right) \right) + A_{k+1} f\left(y^k\right) - A_k \varepsilon$$

$$+ \left\langle \ \nabla f\left(y^{k+1}\right), \alpha_{k+1} \cdot \left(u^{k+1} - y^{k+1}\right) + A_k \cdot \left(x^k - y^{k+1}\right) \ \right\rangle + \frac{\left(1 + A_k \tilde{\mu}\right)}{2} \left\| u^{k+1} - u^k \right\|^2.$$

Следовательно,

$$\varphi_{k+1}^{*} + A_k \varepsilon \geq A_{k+1} \left[f\left(y^{k+1}\right) + \left\langle \nabla f\left(y^{k+1}\right), x^{k+1} - y^{k+1} \right\rangle + \frac{A_{k+1} \cdot \left(1 + A_k \tilde{\mu}\right)}{2\alpha_{k+1}^2} \left\| x^{k+1} - y^{k+1} \right\|^2 + h\left(x^{k+1}\right) \right].$$

Отсюда и из условия (27), $A_{k+1} \cdot \left(1 + A_k \tilde{\mu}\right) / \alpha_{k+1}^2 = L_{k+1}^{j_{k+1}}$ (с учетом (28)) получаем $\varphi_{k+1}^{*} + A_{k+1}\varepsilon = \varphi_{k+1}^{*} + A_k \varepsilon + \alpha_{k+1}\varepsilon \geq A_{k+1} F\left(x^{k+1}\right)$, $L_{k+1}^{j_{k+1}} \leq 2L$. ■

В действительности, выше установлено более сильное утверждение – в оценке (29) можно улучшить константу L.

Теорема 8. *Пусть справедливо предположение 2. Тогда АМТ с (27) для задачи (1) сходится согласно оценке*

$$F\left(x^N\right)-\min_{x\in Q}F\left(x\right)\le\frac{R^2}{A_N}+\varepsilon\le\min\left\{\frac{4LR^2}{\left(N+1\right)^2},LR^2\exp\left(-\frac{N}{2}\sqrt{\frac{\mu}{L\omega_n}}\right)\right\}+\varepsilon,\quad(31)$$

где $L=\max\limits_{k=0,\ldots,N}L_k^{j_k}$. *При этом среднее число вычислений значения функции на одной итерации будет* ≈ 4, *а градиента функции* ≈ 2.

Попробуем сыграть на условии (27), искусственно вводя неточность.

Лемма 5. *Пусть*

$$\left\|\nabla f\left(y\right)-\nabla f\left(x\right)\right\|_{*}\leq L_{\nu}\left\|y-x\right\|^{\nu} \tag{32}$$

при некотором $\nu\in\left[0,1\right]$. *Тогда*

$$f\left(y\right)+\left\langle \nabla f\left(y\right),x-y\right\rangle+\frac{L}{2}\left\|x-y\right\|^{2}+\delta\geq f\left(x\right),\ L=L_{\nu}\left[\frac{L_{\nu}}{2\delta}\frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}}. \tag{33}$$

Основным результатом данного пункта является описание и последующая оценка скорости сходимости нового варианта универсального метода [6] на базе МТ (УМТ).

Универсальный Метод Треугольника

1. $L_{k+1}^{0} := L_{k}^{j_k} / 2,\ j_{k+1} = 0,$

$$u^k = \arg\min_{x \in Q} \varphi_k(x).$$

2. $\alpha_{k+1} := \frac{1 + A_k \tilde{\mu}}{2 L_{k+1}^{j_{k+1}}} + \sqrt{\frac{1 + A_k \tilde{\mu}}{4 \left(L_{k+1}^{j_{k+1}} \right)^2} + \frac{A_k \cdot (1 + A_k \tilde{\mu})}{L_{k+1}^{j_{k+1}}}},\ A_{k+1} := A_k + \alpha_{k+1},$

$$x^{k+1} := \frac{\alpha_{k+1} u^{k+1} + A_k x^k}{A_{k+1}},\ y^{k+1} := \frac{\alpha_{k+1} u^k + A_k x^k}{A_{k+1}}.$$

До тех пор пока

$$f\left(y^{k+1}\right)+\left\langle \nabla f\left(y^{k+1}\right), x^{k+1}-y^{k+1}\right\rangle+\frac{L_{k+1}^{j_{k+1}}}{2}\left\|x^{k+1}-y^{k+1}\right\|^2+\frac{\alpha_{k+1}}{2A_{k+1}}\varepsilon < f\left(x^{k+1}\right),$$

выполнять

$$j_{k+1} := j_{k+1}+1;\ L_{k+1}^{j_{k+1}} := 2^{j_{k+1}} L_k^{j_k}.$$

3. Если не выполнен критерий останова, то $k := k+1$ и **go to** 1.

Теорема 9. *Пусть выполняется условие (32) хотя бы для* $\nu = 0$, *и справедливо предположение 2 с* $\mu \geq 0$ *(допускается брать* $\mu = 0$*). Тогда УМТ для задачи (1) сходится согласно оценке*

$$F\left(x^N\right) - \min_{x \in Q} F\left(x\right) \leq \varepsilon,$$

$$N \approx \min\left\{ \inf_{\nu \in [0,1]} \left(\frac{L_\nu \cdot \left(16R\right)^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+3\nu}}, \inf_{\nu \in [0,1]} \left\{ \left(\frac{8L_\nu^{\frac{2}{1+\nu}} \omega_n}{\mu \varepsilon^{\frac{1-\nu}{1+\nu}}} \right)^{\frac{1+\nu}{1+3\nu}} \ln^{\frac{2+2\nu}{1+3\nu}} \left(\frac{16 L_\nu^{\frac{4+6\nu}{1+\nu}} R^2}{\left(\mu/\omega_n\right)^{\frac{1+\nu}{1+3\nu}} \varepsilon^{\frac{5+7\nu}{2+6\nu}}} \right) \right\} \right\}. \quad (34)$$

При этом среднее число вычислений значения функции на одной итерации будет ≈ 4, *а градиента функции* ≈ 2.

Доказательство. Рассмотрим два случая, когда $\mu \geq 0$ – мало: $\mu \ll \varepsilon / \left(2R^2\right)$, μ – велико: $\mu \gg \varepsilon / \left(2R^2\right)$, см. формулу (23).

В первом случае будем считать, что

$$A_{k+1} / \alpha_{k+1}^2 \approx A_{k+1} \cdot \left(1 + A_k \tilde{\mu}\right) / \alpha_{k+1}^2 = L_{k+1}^{j_{k+1}}, \text{ т.е. } \varepsilon \frac{\alpha_{k+1}}{2A_{k+1}} \approx \frac{\varepsilon}{2} \sqrt{\frac{1}{L_{k+1}^{j_{k+1}} A_{k+1}}}, \quad (35)$$

а во втором случае

$$A_{k+1}^2 \tilde{\mu} / \alpha_{k+1}^2 \approx A_{k+1} \cdot \left(1 + A_k \tilde{\mu}\right) / \alpha_{k+1}^2 = L_{k+1}^{j_{k+1}}, \text{ т.е. } \varepsilon \frac{\alpha_{k+1}}{2A_{k+1}} \approx \frac{\varepsilon}{2} \sqrt{\frac{\tilde{\mu}}{L_{k+1}^{j_{k+1}}}}. \quad (36)$$

Из формулы (31) (см. теорему 8) имеем

$$\frac{R^2}{A_N} + \frac{\varepsilon}{2} \approx \varepsilon,$$

т.е. $A_N \approx 2R^2/\varepsilon$, а также ($L = \max\limits_{k=0,\ldots,N} L_k^{j_k}$)

$$N^2 \approx \frac{8LR^2}{\varepsilon}, \text{ (в первом случае)} \qquad (37)$$

$$N^2 \approx 4\frac{L}{\tilde{\mu}} \ln^2\left(\frac{2LR^2}{\varepsilon}\right). \text{ (во втором случае)} \qquad (38)$$

Из формул (33), (35) – (38) имеем, что в первом случае

$$L \le 2L_\nu \left[\frac{L_\nu}{2\frac{\varepsilon}{2}\sqrt{\frac{1}{LA_N}}} \frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}} \le 2L_\nu \left[\frac{L_\nu \sqrt{LA_N}}{\varepsilon} \frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}} \le 2L_\nu^{\frac{2}{1+\nu}} \left[\frac{N}{2\varepsilon} \frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}}, (39)$$

а во втором случае

$$L \le 2L_\nu \left[\frac{L_\nu}{2\frac{\varepsilon}{2}\sqrt{\frac{\tilde{\mu}}{L}}} \frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}} \le 2L_\nu \left[\frac{L_\nu \sqrt{L/\tilde{\mu}}}{\varepsilon} \frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}} \le 2L_\nu^{\frac{2}{1+\nu}} \left[\frac{N}{2\varepsilon} \frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}}. (40)$$

Подставляя (39) в (37), а (40) в (38) и учитывая, что параметр $\nu \in [0,1]$ можно выбирать произвольно (допускается, что $L_\nu = \infty$ при некоторых ν – важно, чтобы существовало хотя бы одно значение ν при котором $L_\nu < \infty$; по условию $L_0 < \infty$), получим соответственно,

$$N^2 \approx \frac{16 L_\nu^{\frac{2}{1+\nu}} \left[\frac{N}{2\varepsilon} \frac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}} R^2}{\varepsilon} \Rightarrow N^{\frac{1+3\nu}{1+\nu}} \approx \frac{16 L_\nu^{\frac{2}{1+\nu}} R^2}{\varepsilon^{\frac{2}{1+\nu}}}$$

$$\Rightarrow N \approx \inf_{\nu \in [0,1]} \left(\frac{L_\nu \cdot (16R)^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+3\nu}}, \qquad (41)$$

$$N^2 \approx \frac{8L_\nu^{\frac{2}{1+\nu}}\left[\frac{N}{2\varepsilon}\frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}}}{\tilde{\mu}}\ln^2\left(\frac{2L_\nu^2R^2N}{\varepsilon^{3/2}}\right)\Rightarrow$$

$$\Rightarrow N^{\frac{1+3\nu}{1+\nu}} \approx \frac{8L_\nu^{\frac{2}{1+\nu}}}{\tilde{\mu}\varepsilon^{\frac{1-\nu}{1+\nu}}}\ln^2\left(\frac{2L_\nu^2R^2N}{\varepsilon^{3/2}}\right)\Rightarrow N \approx \left(\frac{8L_\nu^{\frac{2}{1+\nu}}}{\tilde{\mu}\varepsilon^{\frac{1-\nu}{1+\nu}}}\right)^{\frac{1+\nu}{1+3\nu}}\ln^{\frac{2+2\nu}{1+3\nu}}\left(\frac{2L_\nu^2R^2N}{\varepsilon^{3/2}}\right)\Rightarrow$$

$$N \approx \inf_{\nu\in[0,1]}\left\{\left(\frac{8L_\nu^{\frac{2}{1+\nu}}}{\tilde{\mu}\varepsilon^{\frac{1-\nu}{1+\nu}}}\right)^{\frac{1+\nu}{1+3\nu}}\ln^{\frac{2+2\nu}{1+3\nu}}\left(\frac{16L_\nu^{\frac{4+6\nu}{1+\nu}}R^2}{\tilde{\mu}^{\frac{1+\nu}{1+3\nu}}\varepsilon^{\frac{5+7\nu}{2+6\nu}}}\right)\right\}. \blacksquare \qquad (42)$$

Универсальный метод треугольника для задач стохастической композитной оптимизации

Предположим теперь, что вместо настоящих градиентов нам доступны только стохастические градиенты $\nabla f(x) \to \nabla f(x, \xi)$.

Предположение 3. *Для всех* $x \in Q$

$$E_\xi \left[\nabla f(x, \xi) \right] = \nabla f(x) \text{ и } E_\xi \left[\left\| f(x, \xi) - \nabla f(x) \right\|_*^2 \right] \le D. \qquad (43)$$

В ряде приложений бывает полезно рассматривать модификацию условия (43)

$$LE_\xi \left[\max_{x, y \in Q} \left\{ \left\langle \nabla f(y, \xi) - \nabla f(y), x - y \right\rangle - \frac{L}{2} \left\| x - y \right\|^2 \right\} \right] \le \tilde{D}.$$

Далее приводится стохастический вариант УМТ (СУМТ). По-видимому, это первая попытка перенесения универсального метода на задачи стохастической оптимизации.

Предварительно введём обозначения

$$\bar{\nabla}^m f(x) = \frac{1}{m}\sum_{k=1}^{m} \nabla f\left(x, \xi^k\right), \tag{44}$$

где ξ^k – независимые одинаково распределенные (так же как ξ) случайные величины. В принципе, можно было бы аналогично ввести

$$\bar{f}^{\tilde{m}}(x) = \frac{1}{\tilde{m}}\sum_{k=1}^{\tilde{m}} f\left(x, \xi^k\right)$$

и распространить приведенные далее результаты на случай, когда и значение функции $f(x)$ необходимо оценивать.

Переопределим последовательность (14)

$$\varphi_0(x) = V(x, y^0) + \alpha_0 \left[f(y^0) + \left\langle \bar{\nabla}^m f(y^0), x - y^0 \right\rangle + \tilde{\mu} V(x, y^0) + h(x) \right],$$

$$\varphi_{k+1}(x) = \varphi_k(x) + \alpha_{k+1} \left[f(y^{k+1}) + \left\langle \bar{\nabla}^m f(y^{k+1}), x - y^{k+1} \right\rangle + \tilde{\mu} V(x, y^k) + h(x) \right].$$

Стохастический Универсальный Метод Треугольника

1. $L_{k+1}^0 := L_k^{j_k} / 2, \; j_{k+1} = 0,$

$$u^k = \arg\min_{x \in Q} \varphi_k(x).$$

2. $$\alpha_{k+1} := \frac{1 + A_k \tilde{\mu}}{2 L_{k+1}^{j_{k+1}}} + \sqrt{\frac{1 + A_k \tilde{\mu}}{4 \left(L_{k+1}^{j_{k+1}} \right)^2} + \frac{A_k \cdot (1 + A_k \tilde{\mu})}{L_{k+1}^{j_{k+1}}}}, \; A_{k+1} := A_k + \alpha_{k+1},$$

$$x^{k+1} := \frac{\alpha_{k+1} u^{k+1} + A_k x^k}{A_{k+1}},\ y^{k+1} := \frac{\alpha_{k+1} u^k + A_k x^k}{A_{k+1}},\ m_{k+1} := \frac{2DA_{k+1}}{L_{k+1}^{j_{k+1}} \alpha_{k+1} \varepsilon}.$$

До тех пор пока

$$f\left(y^{k+1}\right) + \left\langle \bar{\nabla}^{m_{k+1}} f\left(y^{k+1}\right), x^{k+1} - y^{k+1} \right\rangle + \frac{L_{k+1}^{j_{k+1}}}{2} \left\| x^{k+1} - y^{k+1} \right\|^2 + \frac{3\alpha_{k+1}}{2A_{k+1}} \varepsilon < f\left(x^{k+1}\right),$$

выполнять

$$j_{k+1} := j_{k+1} + 1;\ L_{k+1}^{j_{k+1}} := 2^{j_{k+1}} L_k^{j_k}.$$

3. Если не выполнен критерий останова, то $k := k+1$ и **go to** 1.

Теорема 10. *Пусть выполняется условие (32) хотя бы для* $\nu = 0$, *справедливо предположение 2 с* $\mu \geq 0$ *(допускается брать* $\mu = 0$*), справедливо предположение 3. Тогда СУМТ для задачи (1) сходится согласно оценке*

$$E\left[F\left(x^N\right)\right] - \min_{x \in Q} F(x) \leq 2\varepsilon,$$

$$N \approx \min\left\{ \inf_{\nu \in [0,1]} \left(\frac{L_\nu \cdot (32R)^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+3\nu}}, \inf_{\nu \in [0,1]} \left\{ \left(\frac{16 L_\nu^{\frac{2}{1+\nu}} \omega_n}{\mu \varepsilon^{\frac{1-\nu}{1+\nu}}} \right)^{\frac{1+\nu}{1+3\nu}} \ln^{\frac{2+2\nu}{1+3\nu}} \left(\frac{32 L_\nu^{\frac{4+6\nu}{1+\nu}} R^2}{(\mu/\omega_n)^{\frac{1+\nu}{1+3\nu}} \varepsilon^{\frac{5+7\nu}{2+6\nu}}} \right) \right\} \right\}. \quad (45)$$

Оценка (45) – это оценка числа итераций. При этом среднее число вычислений значения функции на одной итерации будет ≈ 4. *Однако не менее интересна оценка числа обращений за стохастическим градиентом*

$$Q \approx 2\min\left\{\frac{4DR^2}{\varepsilon^2}, \frac{2D\omega_n}{\mu\varepsilon}\ln\left(\frac{2L_0^0R^2}{\varepsilon}\right)\right\} + 2N. \qquad (46)$$

Доказательство. По неравенству Фенхеля

$$\left\langle \bar{\nabla}^{m_{k+1}} f\left(y^{k+1}\right) - \nabla f\left(y^{k+1}\right), x^{k+1} - y^{k+1} \right\rangle - \frac{L_{k+1}^{j_{k+1}} / 2}{2} \left\| x^{k+1} - y^{k+1} \right\|^2 \leq$$

$$\leq \frac{2}{L_{k+1}^{j_{k+1}}} \left\| \bar{\nabla}^{m_{k+1}} f\left(y^{k+1}\right) - \nabla f\left(y^{k+1}\right) \right\|_*^2.$$

Поэтому ключевое неравенство (27)

$$f\left(y^{k+1}\right) + \left\langle \bar{\nabla}^{m_{k+1}} f\left(y^{k+1}\right), x^{k+1} - y^{k+1} \right\rangle + \frac{L_{k+1}^{j_{k+1}}}{2} \left\| x^{k+1} - y^{k+1} \right\|^2 + \frac{\alpha_{k+1}}{A_{k+1}} \frac{\varepsilon}{2} \geq f\left(x^{k+1}\right)$$

переписывается следующим образом

$$f\left(y^{k+1}\right) + \left\langle \nabla f\left(y^{k+1}\right), x^{k+1} - y^{k+1} \right\rangle +$$

$$+\frac{L_{k+1}^{j_{k+1}}/2}{2}\left\|x^{k+1}-y^{k+1}\right\|^2+\frac{\alpha_{k+1}}{A_{k+1}}\frac{\varepsilon}{2}+\frac{\alpha_{k+1}}{A_{k+1}}\varepsilon\geq f\left(x^{k+1}\right). \qquad (47)$$

При получении неравенства (47) для большей наглядности заменяем в рассуждениях правую часть в неравенстве Фенхеля оценкой его математического ожидания, равной согласно условию (43)

$$\frac{2D}{L_{k+1}^{j_{k+1}}m_{k+1}}=\frac{\alpha_{k+1}}{A_{k+1}}\varepsilon.$$

В действительности, тут требуются более громоздкие рассуждения, которые приведут к необходимости увеличения в несколько раз (во сколько именно раз, зависит от “тяжести” хвостов распределения $\nabla f\left(x,\xi\right)$ и от выбранного доверительного уровня) константы 2 в

формуле выбора m_{k+1} в описании СУМТ, и к соответствующему увеличению N и Q.

Оценим число обращений за стохастическим градиентом Q, используя схему доказательства теоремы 9. Для этого, прежде всего, заметим, что $A_N \approx 2R^2/\varepsilon$. Рассмотрим два случая, когда $\mu \geq 0$ – мало: $\mu \ll \varepsilon/\left(2R^2\right)$, μ – велико: $\mu \gg \varepsilon/\left(2R^2\right)$.

В первом случае будем считать, что (см. формулу (35))

$$A_{k+1}/\alpha_{k+1}^2 \approx A_{k+1}\cdot\left(1+A_k\tilde{\mu}\right)/\alpha_{k+1}^2 = L_{k+1}^{j_{k+1}}.$$

а во втором случае, что (см. формулу (36))

$$A_{k+1}^2\tilde{\mu}/\alpha_{k+1}^2 \approx A_{k+1}\cdot\left(1+A_k\tilde{\mu}\right)/\alpha_{k+1}^2 = L_{k+1}^{j_{k+1}}.$$

В первом случае число обращений за стохастическим градиентом оценивается, соответственно, как

$$Q \approx \sum_{k=0}^{N} \frac{2DA_k}{L_k^{j_k}\alpha_k\varepsilon} = \frac{2D}{\varepsilon}\sum_{k=0}^{N}\frac{A_k}{L_k^{j_k}\alpha_k} \approx \frac{2D}{\varepsilon}\sum_{k=0}^{N}\alpha_k = \frac{2D}{\varepsilon}A_N \approx \frac{4DR^2}{\varepsilon^2}, \qquad (48)$$

$$Q \approx \sum_{k=0}^{N} \frac{2DA_k}{L_k^{j_k}\alpha_k\varepsilon} = \frac{2D}{\varepsilon}\sum_{k=0}^{N}\frac{A_k}{L_k^{j_k}\alpha_k} \approx \frac{2D}{\tilde{\mu}\varepsilon}\sum_{k=0}^{N}\frac{\alpha_k}{A_k} \approx \frac{2D}{\tilde{\mu}\varepsilon}\int_1^{A_N/\alpha_0}\frac{dA}{A} \approx$$

$$\approx \frac{2D}{\tilde{\mu}\varepsilon}\ln\left(\frac{A_N}{\alpha_0}\right) \approx \frac{2D}{\tilde{\mu}\varepsilon}\ln\left(\frac{2L_0^0R^2}{\varepsilon}\right). \qquad (49)$$

Из формул (48), (49) получаем оценку (46). Оценка (46) с точностью до логарифмических множителей соответствует нижней оценки. ■

Литература

1. *Nesterov Y.* Smooth minimization of non-smooth function // Math. Program. Ser. A. 2005. V. 103. № 1. P. 127–152.
2. *Nesterov Yu.* Gradient methods for minimizing composite functions // Math. Prog. 2013. V. 140. № 1. P. 125–161.
3. *Devolder O.* Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. CORE UCL, PhD thesis, March 2013.
4. *Devolder O., Glineur F., Nesterov Yu.* First order methods of smooth convex optimization with inexact oracle // Math. Progr. Ser. A. 2014. V. 146 (1-2). P. 37–75.
5. *Devolder O., Glineur F., Nesterov Yu.* Intermediate gradient methods for smooth convex problems with inexact oracle // CORE Discussion Paper 2013/17. 2013.
6. *Nesterov Yu.* Universal gradient methods for convex optimization problems // Math. Prog. 2015. V.152. №1-2.P. 381–404; CORE Discussion Paper 2013/63. 2013.
7. *Devolder O., Glineur F., Nesterov Yu.* First order methods with inexact oracle: the smooth strongly convex case // CORE Discussion Paper 2013/16. 2013.
8. *Гасников А.В., Двуреченский П.Е.* Стохастический промежуточный метод для задач выпуклой оптимизации // ДАН РАН. 2016. Т. 467. № 2. С. 131–134. arXiv:1411.2876
9. *Гасников А.В., Двуреченский П.Е., Нестеров Ю.Е.* Стохастические градиентные методы с неточным оракулом // Труды МФТИ. 2016. Т. 8. № 1. С. 41–91. arxiv:1411.4218

10. *Гасников А.В., Камзолов Д.И., Мендель М.А.* Основные конструкции над алгоритмами выпуклой оптимизации и их приложения к получению новых оценок для сильно выпуклых задач // Труды МФТИ. 2016. Т. 8. № 3. (в печати) arXiv:1603.07701
11. *Nemirovski A.* Lectures on modern convex optimization analysis, algorithms, and engineering applications. Philadelphia: SIAM, 2013.
12. *Nesterov Y.* Primal-dual subgradient methods for convex problems // Math. Program. Ser. B. 2009. V. 120(1). P. 261–283.
13. *Аникин А.С., Гасников А.В., Двуреченский П.Е., Тюрин А.И., Чернов А.В.* Двойственные подходы к задачам минимизации сильно выпуклых функционалов простой структуры при аффинных ограничениях // ЖВМ и МФ. 2016. Т. 56. (подана) arXiv:1602.01686
14. *Juditsky A., Nesterov Yu.* Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization // Stoch. System. 2014. V. 4. no. 1. P. 44–80. arXiv:1401.1792
15. *Немировский А.С., Юдин Д.Б.* Сложность задач и эффективность методов оптимизации. М.: Наука, 1979.
16. *Гасников А.В., Гасникова Е.В., Нестеров Ю.Е., Чернов А.В.* Об эффективных численных методах решения задач энтропийно-линейного программирования // ЖВМ и МФ. 2016. Т. 56. № 4. С. 523–534. arXiv:1410.7719
17. *Аникин А.С., Гасников А.В., Горнов А.Ю.* О неускоренных эффективных методах решения разреженных задач квадратичной оптимизации // Труды МФТИ. 2016. Т. 8. № 2.
18. *Нестеров Ю.Е.* Введение в выпуклую оптимизацию. М.: МЦНМО, 2010.