А. В. Назин, Алгоритмы инерционного зеркального спуска в выпуклых задачах стохастической оптимизации, *Автомат. и телемех.*, 2018, выпуск 1, 100–112

Автоматика и телемеханика, № 1, 2018

© 2018 г. А.В. НАЗИН, д-р физ.-мат. наук (nazine@ipu.ru)
(Институт проблем управления им. В.А. Трапезникова РАН, Москва)

АЛГОРИТМЫ ИНЕРЦИОННОГО ЗЕРКАЛЬНОГО СПУСКА В ВЫПУКЛЫХ ЗАДАЧАХ СТОХАСТИЧЕСКОЙ ОПТИМИЗАЦИИ[1]

Рассматривается задача минимизации математического ожидания выпуклой функции потерь на заданном выпуклом компакте $X \in \mathbb{R}^N$. Предполагается, что оракул последовательно выдает стохастические субградиенты функции потерь в текущих точках с равномерно ограниченным вторым моментом. Цель состоит в модификации известного метода зеркального спуска, предложенного А.С. Немировским и Д.Б. Юдиным в 1979 г. и обобщающего стандартный градиентный метод. Для начала демонстрируется идея нового так называемого метода инерционного зеркального спуска (ИЗС) на примере детерминированной задачи оптимизации с непрерывным временем. В частности, в евклидовом случае реализуется метод тяжелого шарика; отмечается, что новый метод не использует дополнительного усреднения точек. Далее описывается дискретный алгоритм ИЗС; доказывается теорема о верхней границе на ошибку по целевой функции, т.е. на разницу текущего значения средних потерь и минимума.

Ключевые слова: задачи стохастической оптимизации, выпуклая оптимизация, метод зеркального спуска, метод тяжелого шарика, инерционный зеркальный спуск.

1. Введение

Многие задачи прикладного характера, например, регрессия, статистика, управление и идентификация, машинное обучение и др., формально могут быть сведены к задаче минимизации

$$f(x) \to \min_{x \in X}, \tag{1.1}$$

где априори неизвестная функция $f : X \to \mathbb{R}$ выпуклая, а множество X — выпуклый компакт в $\mathbb{R}^N$; см., например, [1–5], где описаны постановки задач и рекуррентные методы решения. В таких задачах для последовательного оценивания точки минимума

$$x^* \in \operatorname*{Argmin}_{x \in X} f(x)$$

предполагается, что на каждом шаге $t = 1, 2, \dots$ имеется возможность наблюдать (или вычислять) субградиент $g_t = g_t(x_{t-1}) \in \partial f(x_{t-1})$ или его стохастический вариант

$$u_t(x_{t-1}) = g_t + \xi_t$$

[1] Работа поддержана Российским научным фондом (грант №16-11-10015).

в текущей точке $x_{t-1} \in X$, где $\partial f(x)$ — субдифференциал функции f в точке x, а ξ_t — возмущение субградиента[2]. Сказанное предполагает, что минимизируемая функция известна с точностью до принадлежности ее заданному классу $\mathcal{F} \ni f$ выпуклых функций (возможно, с дополнительными свойствами гладкости и др.) и, кроме того, что для возможности решения такой неопределенной задачи в (1.1) предполагается, что в каждый текущий момент времени $t \geqslant 1$ можно обращаться к оракулу (иначе, к вычислительному блоку), подставляя ему на вход точку x_{t-1} и получая на выходе величину стохастического субградиента $u_t(x_{t-1})$.

В [6] показано, что в выпуклых задачах при “правильном” выборе параметров зеркального спуска (ЗС) он оказывается эффективным методом в том смысле, что при каждом $t > 1$ верхняя и нижняя границы ошибки (по целевой функции)

$$f(\hat{x}_t) - \min_{x \in X} f(x) \tag{1.2}$$

совпадают с точностью до абсолютной константы; здесь $\hat{x}_t$ — “окончательная” оценка точки минимума к моменту времени t, основанная на предыдущих наблюдениях субградиентов в полученных точках x_k, $k = 0, 1, \dots, t-1$. Зачастую в качестве такой оценки $\hat{x}_t$ берут среднее арифметическое предыдущих точек, т.е.

$$\hat{x}_t = \frac{1}{t} \sum_{k=0}^{t-1} x_k \, . \tag{1.3}$$

Отметим, что принципиально новым в структуре метода ЗС по сравнению с классическими методами градиентного типа является (явное или неявное) присутствие двух пространств: прямого $E \supset X$ с исходной нормой $\|\cdot\|$ и сопряженного E^* с двойственной нормой $\|\cdot\|_*$; подробнее см. далее в разделе 3 или в [7, раздел 3]. В частном, “евклидовом”, случае, когда $E = E^*$ и обе нормы являются евклидовыми, а множество $X = \mathbb{R}^N$ — все исходное пространство, метод ЗС переходит в субградиентный метод

$$x_t = x_{t-1} - \gamma_t u_t(x_{t-1}). \tag{1.4}$$

Напомним, что введение дополнительного инерционного члена в метод градиента может улучшать свойства сходимости алгоритма. Имеется в виду метод тяжелого шарика, предложенный Б.Т. Поляком в 1964 г. [8] (см., также [9]). Значит, и метод ЗС разумно обобщить посредством добавления соответствующего инерционного члена. Реализации и изучению этой идеи посвящены разделы 4 и 5.

В целом, структура статьи следующая. В разделе 2 формулируется задача стохастической оптимизации и дается формальная постановка. В разделе 3 приводятся предварительные результаты из выпуклого анализа, а также

[2] Речь идет о понятии оракула 1-го порядка в рассматриваемой задаче оптимизации [6], либо в детерминированной постановке, когда $\xi_t \equiv 0$, либо в стохастической при $\mathbb{E}\{\xi_t\} \equiv 0$ и/или других ограничениях на распределение ξ_t.

необходимые определения и термины, используемые при описании и обосновании алгоритмов ЗС. В разделе 4 излагается идея метода инерционного зеркального спуска (ИЗС) в непрерывном времени и проводится предварительное исследование, определяющее соответствующий “импульс силы инерции” и показывающее схему получения верхней границы ошибки (1.2). Далее, в разделе 5 описывается дискретный алгоритм ИЗС, формулируются и доказываются основные результаты. Наконец, приводятся заключение, приложение и список литературы.

2. Задача стохастической оптимизации и формальная постановка

Рассмотрим задачу минимизации вида

$$f(x) \triangleq \mathbb{E}\, Q(x, Z) \to \min_{x \in X}, \tag{2.1}$$

где $Q : X \times \mathcal{Z} \to \mathbb{R}_+$ — функция потерь, содержащая случайную величину Z, распределенную на пространстве $\mathcal{Z}$ неизвестным образом, $\mathbb{E}$ — математическое ожидание, множество $X \subset \mathbb{R}^N$ — заданный выпуклый компакт в N-мерном пространстве, случайная функция $Q(\,\cdot\,, Z) : X \to \mathbb{R}_+$ выпуклая почти наверное (п.н.) на X. Следовательно, функция $f(x)$ также выпуклая на X. Предполагаем, что $f(x) < \infty$ на X; других ограничений нет, т.е. минимизируемая функция принадлежит классу $\mathcal{F}$ всех выпуклых ограниченных функций $f : X \to \mathbb{R}_+$.

Пусть к моменту времени $t \in \mathbb{Z}_+$ задана обучающая выборка в виде последовательности независимых одинаково распределенных (н.о.р.) случайных величин $(Z_1, \ldots, Z_{t-1})$, причем все Z_i имеют такое же распределение на $\mathcal{Z}$, что и Z. Введем обозначение стохастических субградиентов

$$u_k(x) = g(x, Z_k), \quad k = 1, 2, \ldots, \tag{2.2}$$

представляющих собой измеримые функции на $X \times \mathcal{Z}$ такие, что для любого $x \in X$ математическое ожидание $\mathbb{E}\, u_i(x)$ принадлежит субдифференциалу функции f в точке x.

Задача состоит в построении и исследовании новых рекуррентных алгоритмов ЗС, предназначенных для минимизации (2.1) и использующих на каждом шаге $t = 1, 2, \ldots$ стохастический субградиент $u_t(x_{t-1})$ (2.2) в текущей точке $x = x_{t-1} \in X$.

3. Предварительные результаты, определения и термины

Приведем предварительные результаты из выпуклого анализа [10, 11], а также необходимые определения и термины, используемые при описании и обосновании алгоритмов ЗС [6, 7, 12].

Обозначим через E пространство $\mathbb{R}^N$, оснащенное исходной нормой $\|x\|$ (не обязательно евклидовой), где $x = (x^{(1)}, \ldots, x^{(N)})^{\mathrm{T}} \in E$, а через E^* — двойственное пространство векторов $z = (z^{(1)}, \ldots, z^{(N)})^{\mathrm{T}} \in \mathbb{R}^N$, оснащенное сопряженной (или двойственной) нормой

$$\|z\|_* = \max_{\|x\|=1} z^{\mathrm{T}} x \quad \forall\, z \in E^*.$$

Пусть X — выпуклое замкнутое подмножество E и пусть заданы параметр $\beta > 0$ и выпуклая функция $V : X \to \mathbb{R}$. Назовем *β-сопряженной* функцией к V преобразование Лежандра–Фенхеля W_β от произведения βV:

$$\forall\, z \in E^* \quad W_\beta(z) = \sup_{x\in X} \left\{ -z^{\mathrm{T}} x - \beta V(x) \right\} . \tag{3.1}$$

Введем теперь ключевое предположение (по сути, это условие Липшица в сопряженных нормах $\|\cdot\|$ и $\|\cdot\|_*$), которое будет использовано при доказательстве теоремы 1.

Предположение (L). Выпуклая функция $V : X \to \mathbb{R}$ такова, что ее β-сопряженная W_β непрерывно дифференцируема на E^* с градиентом ∇W_β, удовлетворяющим неравенству

$$\|\nabla W_\beta(z) - \nabla W_\beta(\,\tilde z\,)\| \leqslant \frac{1}{\alpha\beta}\|z-\tilde z\|_* \quad \forall\, z, \tilde z \in E^*,\ \beta > 0, \tag{3.2}$$

где α — положительная постоянная, не зависящая от β.

Формально неравенство (3.2) означает липшицевость градиента функции W_β при произвольном $\beta > 0$, а при доказательстве (дискретных) алгоритмов ЗС позволяет контролировать “мощность” стохастических субградиентов. Как известно (см., например, [13, 14]), это неравенство связано со следующим понятием сильной выпуклости функции V.

Определение 1. Пусть $\alpha > 0$. Выпуклая функция $V : X \to \mathbb{R}$ называется α-сильно выпуклой *относительно исходной нормы $\|\cdot\|$, если*

$$V(sx + (1-s)y) \leqslant sV(x) + (1-s)V(y) - \frac{\alpha}{2}s(1-s)\|x-y\|^2 \tag{3.3}$$

при любых $x, y \in X$ и $s \in [0, 1]$.

В следующем утверждении приводятся свойства β-сопряженных функций и, в частности, дается достаточное условие для справедливости предположения (L).

Утверждение 1. Пусть функция $V : X \to \mathbb{R}$ выпукла, а параметр β положителен. Тогда β-сопряженная к V функция W_β обладает следующими свойствами:

1. *Функция $W_\beta : E^* \to \mathbb{R}$ выпукла и имеет сопряженную βV, т.е.*

$$\forall\, x \in X \quad \beta V(x) = \sup_{z\in E^*} \left\{ -z^{\mathrm{T}} x - W_\beta(z) \right\} .$$

2. *Если функция V является α-сильно выпуклой относительно нормы $\|\cdot\|$, то*

   (i) *выполнено предположение (L),*

   (ii) $\operatorname*{argmax}_{x\in X} \left\{ -z^{\mathrm{T}} x - \beta V(x) \right\} = -\nabla W_\beta(z) \in X$.

Отметим, что п. 2(ii) утверждения 1 позволяет отображать пространство E^* на множество X, поскольку $[-\nabla W_\beta(z)] \in X$ для любого $z \in E^*$. Доказательство дано в [14, 15].

Определение 2. Назовем функцию $V : X \to \mathbb{R}_+$ прокси-функцией, если она выпукла и

(i) *существует такая точка $x_* \in X$, что* $\min\limits_{x\in X} V(x) = V(x_*)$,

(ii) *выполнено предположение (L).*

Пример 1 [7]. Пусть целое $N \geqslant 2$, параметр $\lambda > 0$, множество $X = X_{N,\lambda}$ представляет собой λ-симплекс

$$X_{N,\lambda} = \left\{ x = (x^{(1)}, \dots, x^{(N)})^{\mathrm{T}} \in \mathbb{R}^N_+ \, : \, \sum_{i=1}^{N} x^{(i)} = \lambda \right\} ,$$

а пространство $E = \ell_1$, т.е. исходная норма $\|x\| = \sum_{j=1}^N |x^{(j)}|$. Рассмотрим прокси-функцию энтропийного типа

$$V(x) = \lambda \ln\left(\frac{N}{\lambda}\right) + \sum_{j=1}^{N} x^{(j)} \ln x^{(j)}, \quad x \in X_{N,\lambda} \tag{3.4}$$

(где $0 \ln 0 \triangleq 0$), имеющую единственную точку минимума $x_* = (\lambda/N, \dots$ $\dots, \lambda/N)^{\mathrm{T}}$, причем $V(x_*) = 0$. Нетрудно проверить (см. приложения 1 и 2 в [7]), что эта функция является α-сильно выпуклой относительно нормы $\|\cdot\|$ с параметром $\alpha = 1/\lambda$. Важная особенность прокси-функции (3.4) состоит в том, что для нее задача оптимизации в (5.5) может быть решена явно. При этом $W_\beta(z)$ и $\nabla W_\beta(z)$ задаются формулами:

$$\forall\, z \in E^* \quad W_\beta(z) = \lambda\beta \, \ln\left(\frac{1}{N}\sum_{k=1}^{N} e^{-z^{(k)}/\beta}\right), \tag{3.5}$$

$$\frac{\partial W_\beta(z)}{\partial z^{(j)}} = -\lambda e^{-z^{(j)}/\beta} \left(\sum_{k=1}^{N} e^{-z^{(k)}/\beta}\right)^{-1}, \; j = 1, \dots, N. \tag{3.6}$$

Справедливость предположения (L) для функции (3.4) может быть проверена и прямыми вычислениями (см. приложение в [7]): по сути, здесь нет необходимости прибегать к утверждению 1.

Заметим, что при $\lambda = 1$ верно следующее:

– прокси-функция (3.4) равна информационному расхождению Кульбака между равномерным распределением на множестве $\{1, \dots, N\}$ и распределением на этом множестве, заданным вероятностями $x^{(j)}$, $j = 1, \dots, N$;

– в силу (3.6), компоненты вектора $-\nabla W_\beta(z)$ задают на координатах вектора z распределение Гиббса, в котором β играет роль параметра температуры.

4. Идея метода инерционного зеркального спуска (ИЗС) в непрерывном времени

Идею метода ИЗС удобнее всего рассмотреть в непрерывном времени $t \geqslant 0$ для детерминированной задачи оптимизации без дополнительных ограничений по аналогии с [6, п. 1.4, с. 97–98]. Пусть $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ — выпуклая дифференцируемая функция с неизвестной точкой минимума $x^* \in \mathrm{Argmin} f(x)$ и ее минимальным значением $f^* = f(x^*)$. Рассмотрим следующий непрерывный алгоритм, обобщающий метод ЗС, а именно:

$$\dot{\zeta} = -\nabla f(x), \quad t \geqslant 0, \tag{4.1}$$

$$\mu_t \dot{x} + x = \nabla W(\zeta). \tag{4.2}$$

Здесь точка над переменной обозначает производную по времени t, начальное значение двойственной (сопряженной) переменной $\zeta \in \mathbb{R}^{\mathbb{N}}$ нулевое, т.е.

$$\zeta(0) = 0, \tag{4.3}$$

а начальное значение исходной переменной

$$x(0) = \nabla W(\zeta(0)) = \nabla W(0). \tag{4.4}$$

Функциональным параметром в (4.2) является выпуклая непрерывно дифференцируемая функция $W : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}_+$, имеющая сопряженную функцию (преобразование Лежандра)

$$V(x) = \sup_{\zeta \in \mathbb{R}^{\mathbb{N}}} \{\langle \zeta, x \rangle - W(\zeta)\}, \tag{4.5}$$

где $\langle \zeta, x \rangle = \zeta^{\mathrm{T}} x$ — скалярное произведение. Без ограничения общности полагаем $W(0) = 0$. Следовательно, функция V выпуклая и $V(0) = 0$; кроме того, $\nabla W(0) = 0$[3].

Замечание 1. При нулевом параметре $\mu_t \equiv 0$ в (4.2) алгоритм (4.1)–(4.2) представляет собой метод ЗС в непрерывном времени [6]; в частности, при тождественном отображении $\nabla W(\zeta) \equiv \zeta$ и $\mu_t \equiv 0$ получаем стандартный метод градиента

$$\dot{x} = -\nabla f(x), \quad t \geqslant 0. \tag{4.6}$$

При $\mu_t \equiv \mu > 0$ и $W(\zeta) \equiv \zeta$ алгоритм (4.1)–(4.2) сводится к непрерывному методу тяжелого шарика (МТШ) [9]

$$\mu \ddot{x} + \dot{x} = -\nabla f(x), \quad t \geqslant 0. \tag{4.7}$$

Наконец, в случае $\mu_t = t$ (который будет особо интересен далее) и при $W(\zeta) \equiv \equiv \zeta$ алгоритм (4.1)–(4.2) переходит в уравнение

$$t\ddot{x} + 2\dot{x} + \nabla f(x) = 0\,, \tag{4.8}$$

которое при замене времени $t \mapsto 2\sqrt{t}$ принимает вид

$$\ddot{x} + \frac{3}{t}\,\dot{x} + \nabla f(x) = 0\,. \tag{4.9}$$

[3] В общем рассмотрении вводят функцию $V(x, x^0) = \sup_{\zeta \in \mathbb{R}^{\mathbb{N}}} \{\langle x - x^0, \zeta \rangle - W(\zeta)\}$, $x \in \mathbb{R}^N$, с параметром $x^0 \in \mathbb{R}^N$. Тогда будет $V(x^0, x^0) = 0$. В нашем изложении (для простоты) параметр $x^0 = 0$ и опущен.

Учитывая результаты из [16], где подробно исследуется уравнение (4.9) как непрерывная модель алгоритма Нестерова метода ускоренного градиента (для достаточно гладких детерминированных задач) [17], укажем на связь развиваемого здесь подхода с алгоритмами типа ускоренного градиента. Планируется изучить и использовать эту связь в контексте метода ИЗС более подробно в отдельной работе. $\square$

Далее для определенности будем рассматривать параметр $\mu_t \geqslant 0$ дифференцируемым по $t \geqslant 0$, а метод (4.1)–(4.4) называть непрерывным методом ИЗС. Отметим, что система дифференциальных уравнений (4.1)–(4.2) с начальными значениями (4.3)–(4.4) может не удовлетворять классическим условиям существования и единственности ее решения. Примером является рассмотренный выше случай $\mu_t = t$, приводящий к уравнению (4.8), в котором старший коэффициент (при $\ddot{x}$) нулевой при $t = 0$. Однако в [16] доказаны существование и единственность решения эквивалентного уравнения (4.9) (получающегося из (4.8) заменой времени $t \mapsto 2\sqrt{t}$) с начальными условиями $x(0) = \dot{x}(0) = 0$. Легко видеть, что в этом случае и метод (4.1)–(4.4) имеет решение $\{x(t), \zeta(t)\}_{t\geqslant 0}$. Обобщение этого результата на более общий случай (4.1)–(4.4) представляет собой самостоятельное исследование и выходит за рамки данной работы. Далее в этом разделе будем предполагать[4], что существует $\{x(t), \zeta(t)\}_{t\geqslant 0}$ — решение системы (4.1)–(4.2) с указанными начальными условиями и параметрами $\{\mu_t\}_{t\geqslant 0}$ и $W(\cdot)$; без ограничения общности полагаем $W(0) = 0$.

Исследуем свойства непрерывного метода ИЗС, используя функцию

$$W_*(\zeta) = W(\zeta) - \langle \zeta, x^* \rangle, \quad \zeta \in \mathbb{R}^{\mathbb{N}}, \tag{4.10}$$

которая в вырожденном случае $\mu_t \equiv 0$ представляет собой функцию Ляпунова для непрерывного МЗС (4.1)–(4.2) (хотя бы для гладких целевых функций f) [6], а в общем случае, как показано далее, позволяет с ее помощью получить верхнюю границу ошибки (1.2).

Производная функции (4.10) в силу системы (4.1)–(4.2) равна

$$\frac{d}{dt} W_*(\zeta(t)) = \langle \dot{\zeta}, \nabla W - x^* \rangle = -\langle \nabla f(x), \mu_t \dot{x} + x - x^* \rangle = \tag{4.11}$$

$$= \langle \nabla f(x), x^* - x \rangle - \mu_t \langle \nabla f(x), \dot{x} \rangle \leqslant$$

$$\leqslant f(x^*) - f(x) - \mu_t \frac{d}{dt}[f(x(t)) - f(x^*)], \tag{4.12}$$

где неравенство вытекает из выпуклости $f(\cdot)$. Теперь, интегрируя обе части этого неравенства на отрезке $[0, t]$, $t > 0$, с учетом $W_*(0) = 0$ и обозначения

[4] Это предположение существования и единственности решения уравнений (4.1)–(4.4) не ограничивает нашего последующего рассмотрения, поскольку исследование данного раздела носит предварительный, идейный характер и позволяет только выяснить структуру метода ИЗС. В последующем для дискретного алгоритма вопросов существования и единственности решения не возникает, а корректность легко обосновывается.

$f^* = f(x^*)$, получаем следующее:

$$\int_0^t [f(x(t)) - f^*]dt \leqslant -W_*(\zeta(t)) - \mu_t[f(x(t)) - f^*]\Big|_0^t + \int_0^t [f(x(t)) - f^*]\dot{\mu}_t dt, \tag{4.13}$$

где два последних члена в правой части получены интегрированием по частям. Учитывая (4.5), продолжаем (4.13):

$$\int_0^t [f(x(t)) - f^*]dt \leqslant V(x^*) - \mu_t[f(x(t)) - f^*] + \tag{4.14}$$

$$+ \mu_0[f(x(0)) - f^*] + \left[\sup_{s>0} \dot{\mu}_s\right] \int_0^t [f(x(t)) - f^*]dt. \tag{4.15}$$

Отсюда, собирая в левой части неравенства члены с интегральной и текущей ошибками целевой функции, получаем

$$\left[1 - \sup_{s>0} \dot{\mu}_s\right] \int_0^t [f(x(t)) - f^*]dt + \mu_t[f(x(t)) - f^*] \leqslant V(x^*) + \tag{4.16}$$

$$+ \mu_0[f(x(0)) - f^*]. \tag{4.17}$$

Видно, что с целью контролировать обе компоненты ошибки в левой части (4.16), интегральную и текущую, разумно ввести условия

$$\mu_t > 0, \ \dot{\mu}_t \leqslant 1 \quad \forall t > 0, \quad \mu_0 = 0\,, \tag{4.18}$$

что приводит к неравенству

$$f(x(t)) - f^* \leqslant \frac{1}{\mu_t} V(x^*)\,. \tag{4.19}$$

Теперь максимизируем μ_t при ограничениях (4.18). Получаем

$$\mu_t = t, \quad t \geqslant 0, \tag{4.20}$$

соответствующий непрерывный алгоритм ИЗС

$$\dot{\zeta} = -\nabla f(x), \quad t \geqslant 0, \quad \zeta(0) = 0, \tag{4.21}$$

$$t\dot{x} + x = \nabla W(\zeta), \tag{4.22}$$

и верхнюю границу

$$f(x(t)) - f^* \leqslant \frac{1}{t} V(x^*) \quad \forall t > 0\,. \tag{4.23}$$

Таким образом, по сравнению с классическим МЗС уравнения (4.21)–(4.22) содержат дополнительное слагаемое в левой части (4.22), определяющее “импульс силы инерции” $t\dot{x}$.

Замечание 2. Использование априорной верхней оценки в (4.23) возможно только при известной границе $\overline{V} \geqslant V(x^*)$. Далее в теореме 1 используется именно это предположение при оптимизации верхней оценки дискретного алгоритма ИЗС.

5. Алгоритм инерционного зеркального спуска. Основные результаты

Вернемся к задаче с дискретным временем $t \in Z_+$, сформулированной в конце раздела 4. Запишем дискретный вариант алгоритма ИЗС (4.21)–(4.22) с использованием стохастических субградиентов (2.2) вместо точных $\nabla f(\cdot)$ (эти градиенты существовали в разделе 4 по предположению дифференцируемости f):

$$\tau_t = \tau_{t-1} + \gamma_t, \quad t \geqslant 1,\ \tau_0 = 0, \tag{5.1}$$

$$\zeta_t = \zeta_{t-1} + \gamma_t u_t(x_{t-1}), \quad \zeta_0 = 0, \quad x_0 = -\nabla W_{\beta_0}(\zeta_0), \tag{5.2}$$

$$\tau_t \frac{x_t - x_{t-1}}{\gamma_{t+1}} + x_t = -\nabla W_{\beta_t}(\zeta_t). \tag{5.3}$$

Здесь параметры алгоритма

$$\gamma_t > 0, \quad \beta_t \geqslant \beta_{t-1} > 0 \quad \forall t \geqslant 1, \tag{5.4}$$

а функция W_β определяется по прокси-функции $V : X \to \mathbb{R}_+$ через преобразование типа Лежандра–Фенхеля, а именно:

$$W_\beta(\zeta) = \sup_{x \in X} \left\{ -\zeta^{\mathrm{T}} x - \beta V(x) \right\}, \quad \zeta \in E^*. \tag{5.5}$$

Заметим, что уравнение (5.3) эквивалентно следующему:

$$x_t = \frac{\tau_t}{\tau_t + \gamma_{t+1}} x_{t-1} - \frac{\gamma_{t+1}}{\tau_t + \gamma_{t+1}} \nabla W_{\beta_t}(\zeta_t). \tag{5.6}$$

А так как векторы $[-\nabla W_{\beta_t}(\zeta_t)] \in X$ при каждом $t \geqslant 0$, то в силу (5.1)–(5.2) получаем $x_t \in X$ по индукции.

Пусть далее последовательности $(\gamma_i)_{i \geqslant 1}$ и $(\beta_i)_{i \geqslant 1}$ имеют вид

$$\gamma_i \equiv 1, \quad \beta_i = \beta_0 \sqrt{i+1}, \quad i = 1, 2, \ldots, \quad \beta_0 > 0. \tag{5.7}$$

Тогда система уравнений (5.1)–(5.3) сводится к алгоритму ИЗС:

$$\zeta_t = \zeta_{t-1} + u_t(x_{t-1}), \quad \zeta_0 = 0, \quad x_0 = -\nabla W_{\beta_0}(\zeta_0), \tag{5.8}$$

$$x_t = x_{t-1} - \frac{1}{t+1} \left(x_{t-1} + \nabla W_{\beta_t}(\zeta_t) \right), \quad t \geqslant 1. \tag{5.9}$$

Теорема 1. *Пусть X — выпуклое замкнутое подмножество в $\mathbb{R}^N$, а функция потерь $Q(\cdot,\cdot)$ удовлетворяет условиям, приведенным в разделе 2, и, кроме того,*

$$\sup_{x \in X} \mathbb{E} \| g(x, Z) \|_*^2 \leqslant L_{X,Q}^2, \tag{5.10}$$

где постоянная $L_{X,Q} \in (0, \infty)$. Пусть V — прокси-функция на X с параметром $\alpha > 0$ из предположения (L) и пусть существует $x^* \in \operatorname{Argmin}_{x \in X} f(x)$. Тогда для любого целого $t \geqslant 1$ оценка x_t, определенная алгоритмом (5.1)–(5.3) *со стохастическими субградиентами (2.2) и последовательностями* $(\gamma_i)_{i\geqslant 1}$ *и* $(\beta_i)_{i\geqslant 0}$ *из (5.4) с произвольным* $\beta_0 > 0$, *удовлетворяет неравенству*

$$\mathbb{E}[f(x_{t-1}) - f(x^*)] \leqslant V(x^*)\frac{\beta_t}{\tau_t} + \frac{1}{\tau_t}\sum_{i=1}^{t}\frac{\gamma_i^2 L_{X,Q}^2}{2\alpha\beta_{i-1}}\,.$$

В частном случае, для алгоритма (5.8), (5.9) с параметрами (5.7) справедливо неравенство

$$\mathbb{E}\,f(x_t) - \min_{x\in X} f(x) \leqslant \left(\beta_0 V(x^*) + \frac{L_{X,Q}^2}{\alpha\beta_0}\right)\frac{\sqrt{t+2}}{t+1}\,.$$

Если, кроме того, $\overline{V}$ *— постоянная, такая что* $V(x^*) \leqslant \overline{V}$, *и при известном* $L_{X,Q}$ *выбираем* $\beta_0 = L_{X,Q}\,(\alpha\,\overline{V}\,)^{-1/2}$, *то*

$$\mathbb{E}\,f(x_t) - \min_{x\in X} f(x) \leqslant 2\,L_{X,Q}\left(\alpha^{-1}\overline{V}\,\right)^{1/2}\frac{\sqrt{t+2}}{t+1}. \tag{5.11}$$

В частности, можно взять $\overline{V} = \max_{x\in X} V(x)$.

Доказательство теоремы 1 приведено в Приложении.

Сравним полученные выше результаты со статьей [18], в которой разработаны новые, так называемые субградиентные методы многократного усреднения, предназначенные для решения негладких детерминированных задач выпуклой оптимизации. Формально эти задачи можно рассматривать как предельные, когда случайная величина Z в (2.1) имеет вырожденное распределение, т.е. $Z = \mathrm{const}$ п.н. Отметим, что одним из результатов [18] является получение методов, гарантирующих наилучшую скорость сходимости для всей последовательности точек обращения к оракулу. Интересно, что в частном случае в [18, см. (34)] получен субградиентный метод "с двойным простым усреднением", который в наших обозначениях (см. (5.8), (5.9) и утверждение 1 раздела 3) имеет вид

$$x_t^+ = \operatorname*{argmin}_{x\in X}\left\{x^{\mathrm{T}}\sum_{k=1}^{t} u_k(x_{k-1}) + \beta_t V(x)\right\}, \tag{5.12}$$

$$x_t = x_{t-1} - \frac{1}{t+1}\left(x_{t-1} - x_t^+\right), \quad t \geqslant 1, \tag{5.13}$$

и по сути совпадает с нашим алгоритмом (5.8), (5.9), а теорема 3.2 из [18] аналогична теореме 1.

6. Заключение

Выше рассмотрена известная выпуклая задача стохастической оптимизации с целью построения и исследования новых рекуррентных алгоритмов ЗС. Изложена идея метода ИЗС в непрерывном времени, позволяющая выяснить структуру метода ИЗС и показать схему получения верхней границы метода (впрочем, без полного обоснования существования и единственности решения). Далее, при переходе к дискретному времени, получен рекуррентный алгоритм ИЗС с конкретными последовательностями длины шага (γ_k) и обобщенной температуры (β_k), приведено доказательство. Оказалось, что новый метод не требует дополнительного усреднения точек обращения к оракулу и имеет ту же верхнюю границу ошибки по целевой функции, что и прежний, эффективный метод ЗС (на рассмотренном классе задач) [6, 7]. Представляется интересным дальнейшее исследование нового подхода для других классов целевых функций и требований к оракулу.

В заключение автор считает своим долгом поблагодарить Б.Т. Поляка за внимание к данной статье и плодотворные ее обсуждения, а также коллег, активно участвовавших на заседании семинара 21 марта 2017 г. “Теория автоматического управления”, проводимого в ИПУ РАН под руководством Б.Т. Поляка. Особые благодарности А.Б. Юдицкому за важные обсуждения и ссылку на статью [18], а также анонимным рецензентам за подробные отзывы.

ПРИЛОЖЕНИЕ

Доказательство теоремы 1. Используем уравнение (5.2) и положим $v_i = u_i(x_{i-1})$. Тогда в силу предположения (L) и известного интегрального представления приращения функции $W_{\beta_{i-1}}(\cdot)$, запишем (см. доказательство предложения 2 в [7])

$$W_{\beta_i}(\zeta_i) \leqslant W_{\beta_{i-1}}(\zeta_{i-1}) + \gamma_i v_i^{\mathrm{T}} \nabla W_{\beta_{i-1}}(\zeta_{i-1}) + \frac{\gamma_i^2 \|v_i\|_*^2}{2\alpha\beta_{i-1}}.$$

В левой части также учтено, что $(\beta_i)_{i\geqslant 1}$ — неубывающая последовательность и что для фиксированного ζ отображение $\beta \mapsto W_\beta(\zeta)$ представляет собой невозрастающую функцию. Теперь учтем уравнение (5.3), получая

$$W_{\beta_i}(\zeta_i) \leqslant W_{\beta_{i-1}}(\zeta_{i-1}) - \gamma_i \left(\tau_{i-1}\,\frac{x_{i-1} - x_{i-2}}{\gamma_i} + x_{i-1}\right)^{\mathrm{T}} v_i + \frac{\gamma_i^2 \|v_i\|_*^2}{2\alpha\beta_{i-1}}.$$

Суммируя, получаем с учетом $W_{\beta_0}(\zeta_0) = W_{\beta_0}(0) = 0$ и $\zeta_t = \sum\limits_{i=1}^{t} \gamma_i v_i$, что для всякого $x \in X$ выполняется

$$\sum_{i=1}^{t} \gamma_i (x_{i-1} - x)^{\mathrm{T}} v_i + \sum_{i=1}^{t} \tau_{i-1}\,(x_{i-1} - x_{i-2})^{\mathrm{T}} v_i \leqslant$$

$$\leqslant -\left[W_{\beta_t}(\zeta_t) + x^{\mathrm{T}}\zeta_t\right] + \sum_{i=1}^{t} \frac{\gamma_i^2 \|v_i\|_*^2}{2\alpha\beta_{i-1}}.$$

Обе суммы в левой части этого неравенства преобразуем с учетом условных математических ожиданий

$$\mathbb{E}[v_i \,|\, x_1, \ldots, x_{i-1}] = \mathbb{E}[g(x_{i-1}, Z) \,|\, x_{i-1}] \triangleq \overline{g}(x_{i-1}) \in \partial f(x_{i-1}).$$

Поэтому полные математические ожидания приведут с учетом (5.5) к неравенству

$$\mathbb{E}\left[\sum_{i=1}^{t}\gamma_i(x_{i-1}-x)^{\mathrm{T}}\overline{g}(x_{i-1})+\sum_{i=1}^{t}\tau_{i-1}\,(x_{i-1}-x_{i-2})^{\mathrm{T}}\overline{g}(x_{i-1})\right]\leqslant$$
$$\leqslant \beta_t V(x)+\sum_{i=1}^{t}\frac{\gamma_i^2 L_{X,Q}^2}{2\alpha\beta_{i-1}}.$$

В последнем члене правой части использовано условие (5.10). Далее, используем выпуклость функции f на X, и левая часть неравенства может только уменьшиться:

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{t}\gamma_i(f(x_{i-1})-f(x))+\sum_{i=1}^{t}\tau_{i-1}\,(f(x_{i-1})-f(x_{i-2}))\right]\leqslant\\ &\qquad\leqslant \beta_t V(x)+\sum_{i=1}^{t}\frac{\gamma_i^2 L_{X,Q}}{2\alpha\beta_{i-1}}.\end{aligned}\tag{П.1}$$

Обозначим $\widetilde{f}(x_i)\triangleq f(x_i)-f(x)$ и используем (5.1). В левой части неравенства

$$\sum_{i=1}^{t}\tau_{i-1}\,(f(x_{i-1})-f(x_{i-2}))=\sum_{i=1}^{t-1}\tau_i\,(\widetilde{f}(x_i)-\widetilde{f}(x_{i-1}))=$$
$$=\sum_{i=1}^{t-1}(\tau_i\,\widetilde{f}(x_i)-(\tau_{i-1}+\gamma_i)\widetilde{f}(x_{i-1}))=$$
$$=\tau_{t-1}\,(f(x_{t-1})-f(x))-\sum_{i=1}^{t-1}\gamma_i(f(x_{i-1})-f(x)).$$

Подставляем этот результат в левой части неравенства (П.1) с использованием уравнения (5.1) и делим обе части на τ_t; при $x=x^*$ получаем

$$\mathbb{E}[f(x_{t-1})-f(x^*)]\leqslant V(x^*)\frac{\beta_t}{\tau_t}+\frac{1}{\tau_t}\sum_{i=1}^{t}\frac{\gamma_i^2 L_{X,Q}^2}{2\alpha\beta_{i-1}}.\tag{П.2}$$

Подставляя (5.7) в (П.2), приходим к требуемым неравенствам. Теорема 1 доказана.

СПИСОК ЛИТЕРАТУРЫ

1. *Boyd S., Vandenberghe L.* Convex Optimization. Cambridge University Press, 2004.
2. *Ljung L.* System Identification, Theory for the User. Upper Saddle River: Prentice Hall, 1999 (2nd red.).
3. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Data Mining, Inference and Prediction. New York: Springer, 2001.

4. *Нестеров Ю.Е.* Введение в выпуклую оптимизацию. М.: МЦНМО, 2010.
5. *Hazan E.* Introduction to Online Convex Optimization // Foundat. Trends® Optim. 2016. V. 2. No. 3–4. P. 157–325.
6. *Немировский А.С., Юдин Д.Б.* Сложность задач и эффективность методов оптимизации. М.: Наука, 1979.
7. *Юдицкий А.Б., Назин А.В, Цыбаков А.Б., Ваятис Н.* Рекуррентное агрегирование оценок методом зеркального спуска с усреднением // Проблемы передачи информации. 2005. Т. 41. № 4. С. 78–96.
8. *Поляк Б.Т.* О некоторых способах ускорения сходимости итерационных методов // Журн. вычисл. матем. и мат. физ. 1964. Т. 4. № 5. С. 791–803.
9. *Поляк Б.Т.* Введение в оптимизацию. М.: Наука, 1983.
10. *Рокафеллар Р.* Выпуклый анализ. М.: Мир, 1973.
11. *Магарил-Ильяев Г.Г., Тихомиров В.М.* Выпуклый анализ и его приложения. М.: Книж. дом "ЛИБРОКОМ", 2011.
12. *Beck A., Teboulle M.* Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization // Oper. Res. Lett. 2003. V. 31. No. 3. P. 167–175.
13. *Ben-Tal A., Margalit T., Nemirovski A.* The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography // SIAM J. Optim. 2001. 12. No.1. P. 79–108.
14. *Rockafellar R.T., Wets R.J.B.* Variational analysis. N.Y.: Springer, 1998.
15. *Ben-Tal A., Nemirovski A.S.* The Conjugate Barrier Mirror Descent Method for Non-Smooth Convex Optimization // MINERVA Optim. Center Report. Haifa: Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, 1999.
16. *Su W., Boyd S., Candes E.* A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights // J. Machine Learning Res. 2016. No. 17 (153). P. 1–43.
17. *Нестеров Ю.Е.* Об одном классе методов безусловной минимизации выпуклой функции, обладающих высокой скоростью сходимости // Журн. вычисл. мат. и матем. физ. 1984. Т. 24. № 7. Р. 1090–1093.
18. *Nesterov Yu., Shikhman V.* Quasi-monotone Subgradient Methods for Nonsmooth Convex Minimization // J. Optim. Theory Appl. 2015. No. 165. P. 917–940.

Статья представлена к публикации членом редколлегии А.И. Кибзуном.

Поступила в редакцию 20.03.2017