ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2019, том 59, № 6, с. 961–971

УДК 517.518.86

ОЦЕНКИ ЗАЗОРА ДВОЙСТВЕННОСТИ ДЛЯ СЛАБЫХ ЧЕБЫШЁВСКИХ ЖАДНЫХ АЛГОРИТМОВ В БАНАХОВЫХ ПРОСТРАНСТВАХ

© 2019 г. С. В. Миронов[1,*], С. П. Сидоров[1,]**

([1] 410012 Саратов, ул. Астраханская, 83, Саратовский государственный университет, Россия)

**e-mail: mironovsv@info.sgu.ru*

***e-mail: sidorovsp@info.sgu.ru*

Поступила в редакцию 13.03.2018 г.
Переработанный вариант 28.01.2019 г.
Принята к публикации 08.02.2019 г.

В статье рассматриваются слабые жадные алгоритмы для нахождения разреженных решений задач выпуклой оптимизации в банаховых пространствах. Мы рассматриваем понятие зазора двойственности, значения которого неявно вычисляются на шаге выбора направления наискорейшего спуска на каждой итерации жадного алгоритма. Мы показываем, что эти значения дают верхние оценки разности между значениями целевой функции в текущем состоянии и оптимальной точке. Так как значение целевой функции в оптимальной точке заранее неизвестно, текущие значения зазора двойственности можно использовать, например, в критериях останова жадного алгоритма. В статье мы находим оценки значений зазора двойственности в зависимости от числа итераций для рассматриваемых слабых жадных алгоритмов. Библ. 33.

Ключевые слова: нелинейная оптимизация, жадные алгоритмы, разреженность.

DOI: 10.1134/S0044466919060115

ВВЕДЕНИЕ

Пусть X есть банахово пространство с нормой $\|\cdot\|$. Пусть выпуклая функция E определена на X. Задача выпуклой оптимизации состоит в нахождении приближенного решения следующей задачи:

$$E(x) \to \min_{x \in X}. \tag{1}$$

Следуя идеям работы [1], мы будем рассматривать задачу оптимизации в бесконечномерном пространстве. Это связано с тем, что в большинстве приложений пространство поиска хотя и является конечномерным, его размерность очень велика. Поэтому, так же как и в работе [1], наш интерес связан с получением оценок скорости сходимости алгоритмов, не зависящих от размерности пространства поиска. Очевидно, результаты для бесконечномерного банахова пространства предоставляют такие оценки скорости сходимости.

Многие проблемы в области машинного обучения могут быть сведены к задаче (1) с функцией потерь E [2]. Во многих реальных приложениях необходимо, чтобы оптимальное решение x^* задачи (1) имело простую структуру, например, представляло собой линейную комбинацию *конечного* числа элементов заданного множества (словаря $\mathcal{D}$ в X). Другими словами, x^* должен быть разреженным по отношению к словарю $\mathcal{D}$ в X. Заметим, что требование разреженности можно заменить ограничением на кардинальность (то есть ограничением на количество элементов, используемых в линейных комбинациях элементов словаря $\mathcal{D}$ для построения приближенных решений задачи (1)). Тем не менее во многих случаях задачи оптимизации с ограничением на кардинальность являются задачами неполиномиальной сложности. По этой причине в реальных приложениях предпочитают использовать жадные алгоритмы, которые способны генерировать разреженные решения в силу своего дизайна.

Множество элементов $\mathcal{D}$ пространства X называется *словарем* (см., например, [1]), если каждый элемент $g \in \mathcal{D}$ ограничен по норме единицей, $\|g\| \le 1$, и замыкание линейной оболочки $\mathcal{D}$ совпадает с X, т.е. $\overline{\operatorname{span} \mathcal{D}} = X$. Словарь $\mathcal{D}$ называется симметричным, если $-g \in \mathcal{D}$ для всякого $g \in \mathcal{D}$. Далее мы предполагаем, что словарь $\mathcal{D}$ является симметричным.

Нас интересует задача нахождения решений задачи (1), которые являются разреженными по отношению к словарю $\mathcal{D}$, т.е. мы рассматриваем задачу

$$E(x) \to \inf_{x \in \Sigma_m(\mathcal{D})}, \tag{2}$$

где $\Sigma_m(\mathcal{D})$ есть множество всех m-членных полиномов по отношению к словарю $\mathcal{D}$: $\Sigma_m(\mathcal{D}) = \left\{x \in X : x = \sum_{g \in \Lambda} c_g g,\ \#(\Lambda) = m,\ \Lambda \subset \mathcal{D}\right\}$.

Одним из классов конструктивных методов для нахождения наилучших m-членных приближений является класс жадных алгоритмов. Жадные алгоритмы в силу своей конструкции способны находить разреженные решения по отношению к словарю $\mathcal{D}$. Возможно, метод Франк–Вульфа [3], который также известен как метод условного градиента [4], является одним из самых известных алгоритмов для нахождения оптимальных решений условных задач выпуклой оптимизации. В.Ф. Демьянов и А.М. Рубинов обобщили его на случай произвольных банаховых пространств [5]. Дальнейшее исследование алгоритмов типа Франк–Вульфа можно найти в работах [6]–[8]. В статье [8] приводятся прямые и двойственные результаты по сходимости алгоритмов типа Франк–Вульфа на основе применения идей двойственности, представленных в работе [6]. В последнее время были получены новые результаты по сходимости жадных алгоритмов [9]–[20].

Одной из задач вида (1) является проблема построения регрессии типа LASSO, в которой размер словаря является чрезвычайно большим и может расти экспоненциально по отношению к размерности исходного пространства. Поэтому, чтобы разработать алгоритм, имеющий полиномиальное время работы относительно размерности пространства, необходимо ввести дополнительные структурные ограничения и предположения. В работах [2], [21], [22] предполагается, что можно выполнить линейную оптимизацию по словарю за полиномиальное время по отношению к размерности задачи. Другим ключевым структурным ограничением является то, что целевая функция допускает разреженные приближенные решения. Подобный подход используется в работе [23], в которой применяется метод Франк–Вульфа для поиска равновесного распределения потоков в модели Бэкмана, при этом использование геометрической интерпретации транспортных потоков позволило авторам эффективно снизить размерность пространства поиска. Отметим также статью [24], в которой задача ранжирования объектов по значимости (PageRank) рассматривается как выпуклая задача минимизации над симплексом и, используя разреженную структуру задачи, предложено ее решение с использованием итерационных методов, которые на каждой итерации состоят из подзадач полиномиальной сложности. Необходимо упомянуть работу [25], в которой изучаются проблемы, в которых вспомогательная задача на шаге жадного спуска, являющаяся задачей оптимизации линейной формы на симплексе, может быть эффективно решена (т.е. предполагается наличие так называемого линейного минимизационного оракула).

В алгоритмах, представленных в нашей работе, мы используем развитие этих идей: с одной стороны, на шаге жадного спуска мы вводим ослабляющее условие, которое позволяет для многих задач значительно сократить сложность этой подзадачи. С другой стороны, мы ищем решение в виде разреженной комбинации небольшого числа атомов словаря. Мы изучаем следующие три алгоритма для нахождения решений задачи выпуклой оптимизации, которые являются разреженными по отношению к некоторому словарю, в банаховых пространствах:

- чебышёвский жадный алгоритм (Chebyshev greedy algorithm, CGA),
- слабый чебышёвский жадный алгоритм, (weak Chebyshev greedy algorithm, WCGA),
- слабый чебышёвский жадный алгоритм с ошибкой δ, (weak Chebyshev greedy algorithm with error δ, WCGA(δ)).

Прямой результат о сходимости алгоритма WCGA был получен в [1]. В данной работе мы докажем прямой результат о сходимости алгоритма WCGA(δ). Отметим, что в работе [1] показывается, что алгоритм WCGA для нахождения разреженных решений задачи (2) по отношению к словарю $\mathcal{D}$ также решает и задачу (1).

Развивая идеи [6] и [8], мы вводим понятие зазора двойственности для получения двойственных оценок сходимости алгоритмов WCGA и WCGA(δ) при нахождении разреженных решений

Алгоритм 1. Чебышёвский жадный алгоритм (CGA)

begin

• Положить $G_0 = 0$;

for each $m \geq 1$ **do**

• *(Выбор направления спуска)* Найти элемент словаря $\phi_m \in \mathcal{D}$ такой, что $\langle -E'(G_{m-1}), \varphi_m \rangle = \sup_{s \in \mathcal{D}} \langle -E'(G_{m-1}), s \rangle$;

• *(Чебышёвский поиск)* Найти действительные числа ω_i^* такие, что $E\left(\sum_{i=1}^{m} \omega_i^* \phi_i\right) = \inf_{\omega_i} E\left(\sum_{i=1}^{m} \omega_i \phi_i\right)$;

• *(Переход в новое состояние)* Определить $G_m = \sum_{i=1}^{m} \omega_i^* \phi_i$;

end

задач выпуклой оптимизации типа (2). В статье мы находим оценки значений зазора двойственности в зависимости от числа итераций для рассматриваемых слабых жадных алгоритмов.

1. ЖАДНЫЕ АЛГОРИТМЫ

Для функционала $F \in X^*$ и элемента $f \in X$ в данной статье мы будем использовать следующее обозначение $F(f) = \langle F, f \rangle$.

Пусть $\Omega := \{x \in X : E(x) \leq E(0)\}$ и предположим, что Ω ограничено. Будем предполагать, что функция E дифференцируема по Фреше на Ω. Из свойства выпуклости E следует, что для произвольных $x, y \in \Omega$ справедливо неравенство

$$E(y) \geq E(x) + \langle E'(x), y - x \rangle,$$

где $E'(x)$ есть дифференциал Фреше функции E в точке x.

Мы предполагаем, что существует элемент x^* (не обязательно единственный) банахова пространства X, в котором достигается минимум $E^* := E(x^*)$. Очевидно, что множество всех элементов, в которых достигается минимум, является выпуклым. Отметим, что минимум E^* будет достигаться на элементе (или множестве элементов), принадлежащих множеству Ω.

Пусть $A_1(\mathcal{D})$ означает замыкание (в X) выпуклой оболочки словаря $\mathcal{D}$.

Мы рассматриваем семейство жадных алгоритмов для решения задачи выпуклой оптимизации в банаховом пространстве, которые используют дифференциал Фреше при выборе наискорейшего спуска на каждой итерации.

Алгоритм 1 (CGA) есть итерационный алгоритм, предназначенный для решения задачи (2). Этот алгоритм на каждой итерации $m \geq 1$ переходит к следующему состоянию G_m по индукции, используя текущее состояние G_{m-1} и элемент ϕ_m, полученный на шаге выбора направления наискорейшего спуска жадного алгоритма. На этом шаге максимизируется некоторый линейный функционал, определенный градиентной информацией состояния, полученного на предыдущей итерации алгоритма. Алгоритм 1 является алгоритмом типа алгоритма Франк–Вульфа, так как в каждом текущем состоянии G_{m-1} он использует линеаризацию целевой функции E и осуществляет движение в сторону элемента словаря $\mathcal{D}$, который минимизирует этот линейный функционал.

После того, как найден элемент φ_m на шаге выбора направления наискорейшего спуска жадного алгоритма, осуществить переход в новое состояние G_m из текущего состояния G_{m-1} можно разными способами. В частности, на практике часто используются такие разновидности жадного алгоритма, как градиентный метод, метод приведенного градиента, метод сопряженного градиента, методы преследования (см., например, [3], [26]–[28]). В алгоритмах 1–3 мы используем так называемый поиск типа Чебышёва и выбираем среди всех линейных комбинаций ϕ_i, $i = 1, 2, \ldots, m$, такой элемент G_m из $\operatorname{span}\{\phi_i\}_{i=1}^m$, на котором достигается инфимум функции E.

Алгоритм 2. Слабый чебышёвский жадный алгоритм (WCGA(co))

begin

• Положить $G_0 = 0$;

for each $m \geq 1$ **do**

• *(Выбор направления спуска)* Найти элемент словаря $\phi_m \in \mathcal{D}$ такой, что $\langle -E'(G_{m-1}), \phi_m \rangle \geq t_m \sup_{s \in \mathcal{D}} \langle -E'(G_{m-1}), s \rangle$;

• *(Чебышёвский поиск)* Найти действительные числа ω_i^* такие, что $E\left(\sum_{i=1}^{m} \omega_i^* \phi_i\right) = \inf_{\omega_i} E\left(\sum_{i=1}^{m} \omega_i \phi_i\right)$;

• *(Переход в новое состояние)* Определить $G_m = \sum_{i=1}^{m} \omega_i^* \phi_i$;

end

Алгоритм 3. Слабый чебышёвский жадный алгоритм с ошибкой δ (WCGA(δ))

begin

• Положить $G_0 = 0$ и зафиксировать ошибку $\delta > 0$;

for each $m \geq 1$ **do**

• *(Выбор направления спуска)* Найти такой элемент словаря $\phi_m \in \mathcal{D}$, что $\langle -E'(G_{m-1}), \phi_m \rangle \geq t_m \sup_{s \in \mathcal{D}} \langle -E'(G_{m-1}), s \rangle$;

• *(Чебышёвский поиск)* Найти действительные числа ω_i^* такие, что $E\left(\sum_{i=1}^{m} \omega_i^* \phi_i\right) \leq \inf_{\omega_i} E\left(\sum_{i=1}^{m} \omega_i \phi_i\right) + \delta$;

• *(Переход в новое состояние)* Определить $G_m = \sum_{i=1}^{m} \omega_i^* \phi_i$;

end

Отметим, что на шаге выбора направления спуска жадного алгоритма мы ищем супремум среди элементов словаря $\mathcal{D}$ (а не замыкания его выпуклой оболочки $A_1(\mathcal{D})$), поскольку почти все точки множества $A_1(\mathcal{D})$ есть линейные комбинации бесконечного числа элементов словаря, и в связи с этим оптимальное решение, полученное таким образом, не обязано быть разряженным по отношению к словарю $\mathcal{D}$.

Для многих прикладных задач нахождение точного решения подзадачи $\sup_{s \in \mathcal{D}} \langle -E'(G_{m-1}), s \rangle$ на шаге выбора направления наискорейшего спуска жадного алгоритма может оказаться слишком сложным с вычислительной точки зрения (или даже невозможным). В связи с этим интерес представляет следующая модификация чебышёвского жадного алгоритма CGA с ослаблением требования нахождения оптимального решения [1], [19]. Пусть $\tau := \{t_m\}_{m=1}^{\infty}$ будет последовательностью неотрицательных чисел $t_m \leq 1$, $m = 1, 2, \ldots$, которую мы будем называть *ослабляющей*. Алгоритм 2 использует ослабляющую последовательность τ на шаге выбора направления наискорейшего спуска жадного алгоритма для нахождения приближенного решения ϕ_m этой подзадачи, который имеет качество приближения (мультипликативно) по меньшей мере t_m на шаге m. Именно поэтому алгоритм называется *слабым*.

Далее, на шаге чебышёвского поиска Алгоритма 2 мы пытаемся решить задачу выпуклой оптимизации относительно переменных ω_i, $i = 1, \ldots, m$, с ошибкой δ. В книге [29] приведены быстрые алгоритмы для приближенного решения такого рода задач (см. также [19]).

На шаге чебышёвского поиска Алгоритма 2 предполагается существование оптимальных значений ω_i, $i = 1, \ldots, m$. Однако Алгоритм 3 использует возможность приближенного решения, и данное предположение не требуется.

Алгоритмы используют один атом из словаря на каждой итерации, и поэтому гарантируют разреженность получаемых приближенных решений. Алгоритмы являются слабыми в том смысле, что они решают линейные задачи оптимизации на шаге выбора направления наискорейшего спуска приближенно. Более того, один из рассматриваемых в статье алгоритмов (WCGA(δ)) использует также приближенное решение на шаге чебышёвского поиска.

Статья [30] анализирует другую модификацию алгоритма WCGA – приближенный слабый чебышёвский жадный алгоритм (the approximate weak Chebyshev greedy algorithm, AWCGA). AWCGA проводит чебышёвский поиск с приближенным нахождением n-членного аппроксиманта, однако, в отличие от WCGA(δ), алгоритм AWCGA использует мультипликативный нормирующий функционал.

2. ПРЯМЫЕ ОЦЕНКИ

Модуль гладкости функции E на ограниченном множестве Ω определяется следующим образом:

$$\rho(E,u)=\frac{1}{2}\sup_{x\in\Omega,\|y\|=1}\left|E(x+uy)+E(x-uy)-2E(x)\right|.$$

Функция E называется равномерно гладкой на Ω, если $\lim\limits_{u\to 0}\rho(E,u)/u=0$.

Лемма 1 (см. [1, лемма 1.1]). *Пусть E дифференцируема по Фреше и выпукла на Ω. Тогда для всех $x\in\Omega$*

$$0\le E(x+uy)-E(x)-u\langle E'(x),y\rangle\le 2\rho\left(E,u\|y\|\right).$$

Следующая лемма приводится в [1] как лемма 2.2.

Лемма 2. *Пусть F есть ограниченный линейный функционал и пусть $\mathcal{D}$ есть словарь. Тогда*

$$\sup_{s\in\mathcal{D}}\langle F,s\rangle\sup_{s\in A_1(\mathcal{D})}\langle F,s\rangle.$$

Обозначим

$$E^*:=\inf_{x\in X}E(x)=\inf_{x\in\Omega}E(x),$$

$$\mathcal{L}_M:=\{s\in X: s/M\in A_1(\mathcal{D})\},$$

$$A_\epsilon:=A(E,\epsilon)=\inf\{M:\exists y\in\mathcal{L}_M\ \text{такой, что}\ E(y)-E^*\le\epsilon\},$$

$$A_0:=\inf\{M: x^*\in\mathcal{L}_M\}.$$

Используя геометрические свойства функции E, в статье [1] получена, в частности, следующая оценка скорости сходимости алгоритмов CGA, WCGA.

Теорема 1 (прямая оценка для CGA, WCGA). *Пусть E есть равномерно гладкая выпуклая функция с модулем гладкости $\rho(E,u)\le\gamma u^q$, $1<q\le 2$. Тогда для ослабляющей последовательности $\tau=\{t_m\}_{m=1}^{\infty}$, $0<t_m\le 1$, $m=1,2,\ldots$, для CGA (при $t_m=1$, $m=1,2,\ldots$) и WCGA справедлива оценка*

$$E(G_m)-E^*\le C(E,q,\gamma)A_0^q m^{1-q},\tag{3}$$

где $C(E,q,\gamma)$ есть положительное число, не зависящее от m.

Заметим, что в статье [1] получена более общая оценка

$$E(G_m)-E^*\le C(E,q,\gamma)\epsilon_m,\tag{4}$$

где

$$\epsilon_m:=\inf\{\epsilon: A_\epsilon^q m^{1-q}\le\epsilon\}.\tag{5}$$

Неравенство (3) следует из (4), так как $x^*\in\mathcal{L}_{A_0}$.

Лемма 3. *Пусть неотрицательные числа $a_0,a_1,\ldots,a_N$ таковы, что*

$$a_m\le a_{m-1}+\inf_\lambda(-\lambda v a_{m-1}+B\lambda^q)+\delta,\quad B>0,\quad \delta\in[0,1],$$

для $m \le K := [\delta^{-1/q}]$, $q \in (1,2]$. *Тогда*

$$a_m \le C(q,v,B,a_0)m^{1-q}, \quad m \le K.$$

Лемма 4. *Пусть* E *есть равномерно гладкая выпуклая функция с модулем гладкости* $\rho(E,u)$ *on* Ω. *Пусть* f_ϵ *принадлежит* $\mathscr{L}_{A_\epsilon}$. *Тогда для WCGA*(δ) *будет*

$$E(G_m) \le E(G_{m-1}) + \inf_{\lambda \ge 0}(-\lambda t_m A_\epsilon^{-1}(E(G_{m-1}) - E(f)) + 2\rho(E, C_0\lambda)) + \delta,$$

для $m = 1,2,\ldots,\delta^{-1/q}$.

Доказательство. Из определения G_m в описании алгоритма WCGA(δ) мы имеем

$$G_m = \sum_{i=1}^{m} \omega_i^* \phi_i.$$

Шаг чебышёвского поиска WCGA(δ) влечет

$$E(G_m) \le \inf_{\omega_i} E\left(\sum_{i=1}^{m} \omega_i \phi_i\right) + \delta \le \inf_{\lambda \ge 0, \omega} E(G_{m-1} - \omega G_{m-1} + \lambda \phi_m) + \delta. \tag{6}$$

Из леммы 1 следует, что

$$\begin{aligned} E(G_{m-1} - \omega G_{m-1} + \lambda\phi_m) \le E(G_{m-1}) + \lambda\langle -E'(G_{m-1}), \phi_m\rangle - \\ - \omega\langle -E'(G_{m-1}), G_{m-1}\rangle + 2\rho(E, \|\lambda\phi_m - \omega G_{m-1}\|). \end{aligned} \tag{7}$$

Из шага выбора направления спуска WCGA(δ) имеем

$$\langle -E'(G_{m-1}), \phi_m\rangle \ge t_m \sup_{s \in \mathscr{D}} \langle -E'(G_{m-1}), s\rangle. \tag{8}$$

Из леммы 2 получаем

$$\begin{aligned} t_m \sup_{s\in\mathscr{D}}\langle -E'(G_{m-1}), s\rangle = t_m \sup_{s \in A_1(\mathscr{D})} \langle -E'(G_{m-1}), s\rangle = t_m A_\epsilon^{-1} \sup_{s \in A_1(\mathscr{D})} \langle -E'(G_{m-1}), A_\epsilon s\rangle = \\ = t_m A_\epsilon^{-1} \sup_{f \in \mathscr{L}_{A_\epsilon}} \langle -E'(G_{m-1}), f\rangle \ge t_m A_\epsilon^{-1} \langle -E'(G_{m-1}), f_\epsilon\rangle. \end{aligned} \tag{9}$$

Возьмем $\omega = \lambda t_m A_\epsilon^{-1}$, получим

$$\begin{aligned} E(G_{m-1} - \lambda t_m A_\epsilon^{-1} G_{m-1} + \lambda\phi_m)) \le E(G_{m-1}) - \\ - \lambda t_m A_\epsilon^{-1}\langle -E'(G_{m-1}), f_\epsilon - G_{m-1}\rangle + 2\rho\left(E, \lambda\left\|\phi_m - t_m A_\epsilon^{-1} G_{m-1}\right\|\right). \end{aligned} \tag{10}$$

Из выпуклости E имеем

$$\langle -E'(G_{m-1}), f_\epsilon - G_{m-1}\rangle \ge E(G_{m-1}) - E(f_\epsilon). \tag{11}$$

Из (6)–(11) следует, что

$$E(G_m) \le E(G_{m-1}) + \inf_{\lambda \ge 0}\left(-\lambda t_m A_\epsilon^{-1}(E(G_{m-1}) - E(f_\epsilon)) + 2\rho\left(E, \lambda\left\|\phi_m - t_m A_\epsilon^{-1} G_{m-1}\right\|\right)\right) + \delta.$$

Так как $E(G_{m-1}) \le E(0)$, имеем $G_{m-1} \in \Omega$. Наше предположение об ограниченности множества Ω приводит к тому, что существует константа C_1 такая, что $\|G_{m-1}\| \le C_1$. Так как $\phi_m \in \mathscr{D}$, получаем $\|\phi_m\| \le 1$. Таким образом,

$$\left\|t_m A_\epsilon^{-1} G_{m-1} - \phi_m\right\| \le A_\epsilon^{-1} C_1 + 1 =: C_0,$$

что завершает доказательство леммы.

Теорема 2 (прямая оценка для WCGA(δ)). *Пусть* E *есть равномерно гладкая выпуклая функция с модулем гладкости* $\rho(E,u) \le \gamma u^q$, $1 < q \le 2$. *Тогда, для ослабляющей последовательности* $\tau = \{t_k\}_{k=1}^{\infty}$, $0 < t_k \le 1$, $k = 1,2,\ldots$, *имеем для WCGA*(δ)

$$E(G_m) - E^* \le C(E,q,\gamma)A_0^q m^{1-q}, \quad m \le \delta^{-1/q}, \tag{12}$$

где $C(E,q,\gamma)$ *есть положительное число, не зависящее от* m.

Доказательство. Докажем более общее неравенство

$$E(G_m) - E^* \le C(E,q,\gamma)\epsilon_m, \quad m \le \delta^{-1/q}, \tag{13}$$

где $C(E,q,\gamma)$ есть положительное число, не зависящее от m. Неравенство (12) следует из (13), так как $x^* \in \mathcal{L}_{A_0}$.

Неравенства (6) и (10) влекут неравенство

$$E(G_m) \le E(G_{m-1}) + \delta. \tag{14}$$

Так как $m \le \delta^{-1/q}$, имеем

$$E(G_m) \le E(0) + 1,$$

и поэтому, $G_m \in \Omega_1 := \{x \in X : E(x) \le E(0) + 1\}$ для $m \le \delta^{-1/q}$. Пусть $a_n := E(G_n) - E(f)$. Из леммы 4 следует, что

$$a_m \le a_{m-1} + \inf_{\lambda>0}(-\lambda t_m M_0^{-1} a_{m-1} + 2\gamma(C_0\lambda)^q) + \delta.$$

Если a_m неотрицательны, то мы применяем лемму 3 со значениями $v = t_m M_0^{-1}$ и $B = 2\gamma C_0^q$.

Обозначим $K := [\delta^{-1/q}]$. Пусть n' есть наименьшее из $[1, K]$ таких, что $a_{n'} < 0$. Тогда из (14) и $m \le \delta^{-1/q}$ следует, что $a_m \le Cm^{1-q}$ для всех $n' \le m \le K$.

3. ЗАЗОР ДВОЙСТВЕННОСТИ И ДВОЙСТВЕННЫЕ ОЦЕНКИ

В данном разделе представлены двойственные результаты о сходимости для алгоритмов CGA, WCGA и WGAFR(δ).

Важной проблемой при выполнении вычислений при решении задач оптимизации является создание “сертификата”, который можно использовать для эффективной проверки того, что решение (с заданной точностью) получено. В статье [31] вводится понятие сертификатов в контексте вычислительных задач с выпуклой структурой. Методы, в которых имеются оценки на сертификат точности, относятся к классу прямо-двойственных методов [32]. Следует отметить, что сертификат точности (accuracy certificate), с одной стороны, позволяет обосновать прямо-двойственность исследуемого метода [31], а с другой, сертификат точности является вычислимым и поэтому могут использоваться в качестве оценки близости текущей точки к оптимальному решению. Дальнейшее развитие этих идей можно найти в работе [33]. В данной работе мы предлагаем в качестве такого сертификата “близости” к оптимальному решению использовать зазор двойственности.

Значения зазора двойственности неявно вычисляются на шаге выбора направления наискорейшего спуска на каждой итерации жадного алгоритма. Мы покажем, что эти значения дают верхние оценки ошибки приближения текущего решения к оптимальному, то есть разности между значениями целевой функции в текущем состоянии и оптимальной точке. Так как значение целевой функции в оптимальной точке заранее неизвестно, текущие значения зазора двойственности можно использовать, например, в критериях останова жадного алгоритма. В статье мы находим оценки значений зазора двойственности в зависимости от числа итераций для рассматриваемых слабых жадных алгоритмов.

Определение 1. Определим *зазор двойственности* для состояния $G \in \Omega$ и ошибки ϵ следующим образом:

$$g(G) = g(G,\epsilon) =: A_\epsilon \sup_{s\in\mathcal{D}} \left\langle E'(G), A_\epsilon^{-1} G - s \right\rangle. \tag{15}$$

Полезное свойство зазора двойственности описывается следующим утверждением.

Утверждение 1. *Пусть E есть выпуклая функция, определенная на банаховом пространстве X. Тогда для любого $x \in \Omega$ будет*

$$E(x) - E(x^*) \le g(x,\epsilon) + \epsilon.$$

Доказательство. Пусть x_ϵ таков, что $E(x_\epsilon) - E(x^*) < \epsilon$ и $x_\epsilon/A_\epsilon \in A_1(\mathcal{D})$, т.е. $x_\epsilon \in \mathcal{L}_{A_\epsilon}$. Вначале покажем, что

$$E(x) - E(x_\epsilon) \le g(x).$$

Так как E является выпуклой на X, для всякого $x \in \Omega$ будет

$$E(x_\epsilon) \ge E(x) + \langle E'(x), x_\epsilon - x\rangle. \tag{16}$$

Так как $x_\epsilon \in \mathcal{L}_{A_\epsilon}$, имеем

$$E(x) + \langle E'(x), x_\epsilon - x\rangle \ge E(x) - \sup_{s\in\mathcal{L}_{A_\epsilon}} \langle E'(x), x - s\rangle = E(x) - A_\epsilon \sup_{s\in\mathcal{L}_{A_\epsilon}} \langle E'(x), xA_\epsilon^{-1} - sA_\epsilon^{-1}\rangle. \tag{17}$$

Так как $\mathcal{L}_{A_\epsilon} = A_\epsilon A_1(\mathcal{D})$, получаем

$$\begin{aligned} E(x) - A_\epsilon \sup_{s\in\mathcal{L}_{A_\epsilon}} \langle E'(x), xA_\epsilon^{-1} - sA_\epsilon^{-1}\rangle &= E(x) - A_\epsilon \sup_{s'\in A_1(\mathcal{D})} \langle E'(x), xA_\epsilon^{-1} - s'\rangle = \\ &= E(x) - A_\epsilon \sup_{s'\in\mathcal{D}} \langle E'(x), xA_\epsilon^{-1} - s'\rangle, \end{aligned} \tag{18}$$

где мы использовали утверждение леммы 2.

Тогда из (16)–(18) следует, что

$$E(x) - E(x_\epsilon) \le A_\epsilon \sup_{s'\in\mathcal{D}} \langle E'(x), xA_\epsilon^{-1} - s'\rangle =: g(x),$$

и мы получаем наше утверждение (поскольку $E(x_\epsilon) - E(x^*) < \epsilon$).

Таким образом, практическая применимость зазора двойственности связана с тем фактом, что зазор двойственности $g(x)$ является оценкой ошибки приближения оптимального состояния $E(x^*)$ текущим состоянием $E(x)$.

Справедлив следующий двойственный результат для алгоритмов CGA и WCGA.

Теорема 3. *Пусть E есть равномерно гладкая выпуклая функция, определенная на банаховом пространстве X. Пусть $\rho(E,u)$ есть модуль гладкости функции E и предположим, что $\rho(E,u) \le \gamma u^q$, $1 < q \le 2$. Пусть $\tau = \{t_m\}_{m=1}^\infty$, $0 < \theta < t_k < 1$, $k = 1,2,\ldots$, есть ослабляющая последовательность. Предположим, что CGA или WCGA выполнены для $N > 2$ итераций. Тогда найдется такая итерация $1 \le \tilde{m} \le N$, что*

$$g(G_{\tilde{m}}) \le \beta C(E,q,\gamma) A_0^q N^{1-q},$$

где $\beta > 0$ не зависит от N.

Имеет место следующий двойственный результат для алгоритма WCGA(δ).

Теорема 4. *Пусть E есть равномерно гладкая выпуклая функция, определенная на банаховом пространстве X. Пусть $\rho(E,u)$ есть модуль гладкости функции E и предположим, что $\rho(E,u) \le \gamma u^q$, $1 < q \le 2$. Пусть $\tau = \{t_m\}_{m=1}^\infty$, $0 < \theta < t_k < 1$, $k = 1,2,\ldots$, есть ослабляющая последовательность. Предположим, что алгоритм WCGA(δ) выполнен для $0 < N \le \delta^{-1/q}$ итераций. Тогда найдется такая итерация $1 \le \tilde{m} \le N$, что*

$$g(G_{\tilde{m}}) \le \beta C(E,q,\gamma) A_0^q N^{1-q},$$

где $\beta > 0$ не зависит от N.

Достаточно доказать двойственный результат для алгоритма с ошибкой WCGA(δ) (теорему 4) и тогда теорема 3 будет непосредственно следовать из соответствующего результата для WCGA(δ).

Нам потребуются следующие вспомогательные результаты.

Лемма 5. *Пусть E есть равномерно гладкая выпуклая функция, определенная на банаховом пространстве X. Пусть $\rho(E,u)$ означает модуль гладкости функции E. Тогда для WCGA(δ) справедливо следующее неравенство:*

$$E(G_m) \le E(G_{m-1}) + \inf_{\lambda\ge 0}(-\lambda t_m A_\Omega^{-1} g(G_{m-1}) + 2\rho(E, C_0\lambda)) + \delta, \quad m = 1,2,\ldots,$$

где C_0 не зависит от m.

Доказательство. Из определения G_m в описании алгоритма WCGA(δ) имеем

$$G_m = \sum_{i=1}^{m} \omega_i G_i.$$

Шаг чебышёвского поиска для WCGA(δ) влечет

$$E(G_m) = \inf_{\omega_i} E\left(\sum_{i=1}^{m} \omega_i G_i\right) \le \inf_{\lambda\ge 0,\omega} E(G_{m-1} - \omega G_{m-1} + \lambda\phi_m) + \delta. \tag{19}$$

Из леммы 1 следует, что

$$E(G_{m-1} - \omega G_{m-1} + \lambda\phi_m) \le E(G_{m-1}) + \lambda\left\langle -E'(G_{m-1}), \phi_m\right\rangle - \\ - \omega\left\langle -E'(G_{m-1}), G_{m-1}\right\rangle + 2\rho(E, \|\lambda\phi_m - \omega G_{m-1}\|).$$

Из шага выбора направления наискорейшего спуска алгоритма WCGA(δ) следует

$$\left\langle -E'(G_{m-1}), \phi_m\right\rangle \ge t_m \sup_{s\in\mathcal{D}}\left\langle -E'(G_{m-1}), s\right\rangle. \tag{20}$$

Возьмем $\omega = \lambda t_m A_\epsilon^{-1}$, тогда

$$E(G_{m-1} - \omega G_{m-1} + \lambda\phi_m)) \le E(G_{m-1}) - \\ - \lambda t_m\left\langle -E'(G_{m-1}), \phi_m - G_{m-1}A_\epsilon^{-1}\right\rangle + 2\rho\left(E, \lambda\left\|\phi_m - G_{m-1}A_\epsilon^{-1}\right\|\right). \tag{21}$$

Используя (20) и определение зазора двойственности (15), имеем

$$\left\langle -E'(G_{m-1}), \phi_m - G_{m-1}A_\epsilon^{-1}\right\rangle \ge t_m \sup_{s\in\mathcal{D}}\left(-E'(G_{m-1}), s - G_{m-1}A_\epsilon^{-1}\right) = t_m A_\epsilon^{-1} g(G_{m-1}). \tag{22}$$

Из (19), (21) и (22) следует, что

$$E(G_m) \le E(G_{m-1}) + \inf_{\lambda\ge 0}\left(-\lambda t_m A_\epsilon^{-1} g(G_{m-1}) + 2\rho\left(E, \lambda\left\|\phi_m - t_m A_\epsilon^{-1} G_{m-1}\right\|\right)\right) + \delta.$$

Из $E(G_{m-1}) \le E(0)$ следует, что $G_{m-1} \in \Omega$. Предположение ограниченности Ω приводит к тому, что найдется такая константа C_1, что $\|G_{m-1}\| \le C_1$. Так как $\phi_m \in \mathcal{D}$, имеем $\|\phi_m\| \le 1$. Таким образом,

$$\left\|t_m A_\epsilon^{-1} G_{m-1} - \phi_m\right\| \le A_\epsilon^{-1} C_1 + 1 =: C_0,$$

и лемма доказана.

Лемма 6. *Пусть* $0 < \mu < 1$ *есть некоторое действительное число и пусть* M *является натуральным числом. Тогда*

$$\epsilon_{m_0} \le \mu^{1-q}\epsilon_M,$$

где ϵ_m *определено в* (5) *и* $m_0 = [\mu M]$.

Доказательство. Из (5) следует, что

$$\inf\left\{\epsilon : m^{1-q} < \frac{\left(\frac{\epsilon}{\mu^{1-q}}\right)}{A\left(\frac{\epsilon}{\mu^{1-q}}\right)}\right\} = \mu^{1-q}\epsilon_m.$$

Пусть $\epsilon^* := \frac{\epsilon}{\mu^{1-q}}$, т.е. $\epsilon^*\mu^{1-q} = \epsilon$. Заметим, что $\mu^{1-q} > 1$ и $\frac{\epsilon}{\mu^{1-q}} < \epsilon$. Имеем

$$A\left(\frac{\epsilon}{\mu^{1-q}}\right) \ge A_\Omega. \tag{23}$$

Из (23) следует, что

$$\epsilon_{m_0} = \inf\left\{\epsilon : M^{1-q} < \frac{\epsilon^*}{A_\Omega^q}\right\} \le \inf\left\{\epsilon^* : M^{1-q} < \frac{\epsilon^*}{A_{\epsilon^*}^q}\right\} = \mu^{1-q}\epsilon_M.$$

Доказательство теоремы 4. Мы докажем более общее утверждение, следствием которого является утверждение теоремы: найдется такая итерация $1 \le \tilde{m} \le N$, что

$$g(G_{\tilde{m}}) \le \beta C(E,q,\gamma)\epsilon_N,$$

где $\beta > 0$ не зависит от N. Из (13) следует, что

$$E(G_m) - E^* \le C(E,q,\gamma)\epsilon_m, \quad m \le \delta^{-1/q},$$

$$\epsilon_m := \inf\{\epsilon : A_\epsilon^q m^{1-q} \le \epsilon\}.$$

Предположим, что

$$g(G_m) > \beta C(E,q,\gamma)\epsilon_N \tag{24}$$

для всех $[\mu N] \le m \le N$, $0 < \mu < 1$ (μ фиксировано и будет определено позже).

Из леммы 5 с $\lambda = \epsilon_m$ следует, что

$$E(G_{m+1}) - E^* \le E(G_m) - E^* - \epsilon_m t_m A_\Omega^{-1} g(G_m) + 2\gamma(C_0\epsilon_m)^q + \delta. \tag{25}$$

Используя наше допущение (24), неравенство (25) может быть переписано в виде

$$E(G_{m+1}) - E^* \le E(G_m) - E^* - \epsilon_m t_m A_\Omega^{-1}\beta C(E,q,\gamma)\epsilon_N + 2\gamma(C_0\epsilon_m)^q + \delta. \tag{26}$$

Нам потребуются следующие неравенства:

1) $\theta \le t_k \le 1$, $k = 1, 2, \ldots$;

2) $\epsilon_{m_0} \le \mu^{1-q}\epsilon_N$ (лемма 3);

3) так как $[\mu N] \le m \le N$, имеем $\epsilon_{[\mu N]} \ge \epsilon_m \ge \epsilon_N$. Тогда (26) дает

$$E(G_{m+1}) - E^* \le E(G_m) - E^* - A_\Omega^{-1}\beta\theta C(E,q,\gamma)\epsilon_N^2 + 2\gamma(C_0)^q\mu^{q(1-q)}\epsilon_N^q + \delta.$$

Если мы запишем цепь неравенств для всех $m = m_0, \ldots, N$, $m_0 := [\mu N]$, мы получим

$$E(G_{m+1}) - E^* \le E(G_{m_0}) - E^* - (N - m_0)\epsilon_N\left[A_\Omega^{-1}\beta\theta C(E,q,\gamma)\epsilon_N - 2\gamma(C_0)^q\mu^{q(1-q)}\epsilon_N^{q-1} + \frac{\delta}{\epsilon_N}\right] \le$$

$$\le C(E,q,\gamma)\epsilon_{m_0} - N(1-\mu)\epsilon_N\left[A_\Omega^{-1}\beta\theta C(E,q,\gamma)\epsilon_N - 2\gamma(C_0)^q\mu^{q(1-q)}\epsilon_N^{q-1} + \frac{\delta}{\epsilon_N}\right] \le$$

$$\le C(E,q,\gamma)\mu^{1-q}\epsilon_N - N(1-\mu)\epsilon_N\left[A_\Omega^{-1}\beta\theta C(E,q,\gamma)\epsilon_N - 2\gamma(C_0)^q\mu^{q(1-q)}\epsilon_N^{q-1} + \frac{\delta}{\epsilon_N}\right] =$$

$$= \epsilon_N\left(C(E,q,\gamma)\mu^{1-q} - N(1-\mu)\left[A_\Omega^{-1}\beta\varepsilon_N\theta C(E,q,\gamma) - 2\gamma(C_0)^q\mu^{q(1-q)}\epsilon_N^{q-1} + \frac{\delta}{\epsilon_N}\right]\right).$$

Если взять достаточно большое β,

$$\beta > \frac{\dfrac{C(E,q,\gamma)\mu^{1-q}}{N(1-\mu)} + 2\gamma(C_0)^q\mu^{q(1-q)}\epsilon_N^{q-1} - \dfrac{\delta}{\epsilon_N}}{A_\Omega^{-1}\epsilon_N\theta C(E,q,\gamma)},$$

то мы получим $E(G_m) - E^* < 0$, что невозможно. Параметр μ можно взять таким, что

$$\mu := \arg\min_{0\le\mu\le1}\left(\frac{C(E,q,\gamma)\mu^{1-q}}{N(1-\mu)} + 2\gamma(C_0)^q\mu^{q(1-q)}\epsilon_N^{q-1} - \frac{\delta}{\epsilon_N}\right).$$

ЗАКЛЮЧЕНИЕ

Теоремы 1 и 2 дают оценки прямых ошибок приближения оптимального решения $E(x^*)$ текущим решением $E(G)$ для слабых чебышёвских жадных алгоритмов. Однако, так как для многих прикладных задач оптимальное значение $E(x^*)$ и константа γ в модуле гладкости функции E неизвестны, величина оценки ошибки приближения к оптимальному решению на текущей итерации $E(G) - E(x^*)$ неизвестна. Значение зазора двойственности $g(G)$, определенного в (15) и оценки которого получены в теоремах 3 и 4, вычисляется неявно на шаге выбора направления

наискорейшего спуска на каждой итерации жадных алгоритмов WCGA или WCGA(δ) и является подходящей величиной для оценки прямой ошибки $E(G) - E(x^*)$, поскольку зазор двойственности есть верхняя оценка прямой ошибки (см. утверждение 1).

СПИСОК ЛИТЕРАТУРЫ

1. *Temlyakov V.N.* Greedy approximation in convex optimization // Constr. Approx. 2015. V. 41. № 2. P. 269–296.
2. *Bubeck S.* Convex optimization: Algorithms and complexity // Foundations and Trends in Machine Learning. 2015. V. 8. № 3–4. P. 231–357.
3. *Frank M., Wolfe P.* An algorithm for quadratic programming // Naval Res. Logis. Quart.a. 1956. V. 3. № 1–2. P. 95–110.
4. *Левитин Е.С., Поляк Б.Т.* Методы минимизации при наличии ограничений // ЖВММФ. 1966. Т. 6. № 5. С. 787–823.
5. *Демьянов В.Ф., Рубинов А.М.* Приближенные методы решения экстремальных задач. Л.: Изд-во Ленингр. ун-та, 1968.
6. *Clarkson K.L.* Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm // ACM T. Algorithms. 2010. V. 6. № 4. P. 1–30.
7. *Freund R.M., Grigas P.* New analysis and results for the Frank-Wolfe method // Math. Program. 2016. V. 155. № 1. P. 199–230.
8. *Jaggi M.* Revisiting Frank-Wolfe: Projection-free sparse convex optimization // Proceedings of the 30th International Conference on Machine Learning (ICML-13). 2013. P. 427–435.
9. *Friedman J.* Greedy function approximation: A gradient boosting machine // Ann. Stat. 2001. V. 29. № 5. P. 1189–1232.
10. *Davis G., Mallat S., Avellaneda M.* Adaptive greedy approximation // Constr. Approx. 1997. V. 13. № 1. P. 57–98.
11. *Zhang Z., Shwartz S., Wagner L., Miller W.* A greedy algorithm for aligning DNA sequences // J. Comput. Biol. 2000. V. 7. № 1–2. P. 203–214.
12. *Huber P.J.* Projection pursuit // Ann. Stat. 1985. V. 13. P. 435–525.
13. *Jones L.* On a conjecture of huber concerning the convergence of projection pursuit regression // Ann. Stat. 1987. V. 15. № 2. P. 880–882.
14. *Barron A.R., Cohen A., Dahmen W., DeVore R.A.* Approximation and learning by greedy algorithms // Ann. Stat. 2008. V. 36. № 1. P. 64–94.
15. *DeVore R.A., Temlyakov V.N.* Some remarks on greedy algorithms // Adv. Comput. Math. 1996. V. 5. P. 173–187.
16. *Konyagin S.V., Temlyakov V.N.* A remark on greedy approximation in banach spaces // East J. Approx. 1999. V. 5. № 3. P. 365–379.
17. *Nguyen H., Petrova G.* Greedy strategies for convex optimization // Calcolo. 2016. V. 54. № 1. P. 207–224.
18. *Temlyakov V.N.* Dictionary descent in optimization // Anal. Math. 2016. V. 42. № 1. P. 69–89.
19. *DeVore R.A., Temlyakov V.N.* Convex optimization on banach spaces // Found. Comput. Math. 2016. V. 16. № 2. P. 369–394.
20. *Темляков В.Н.* Сходимость и скорость сходимости некоторых гриди-алгоритмов в выпуклой оптимизации // Тр. МИАН. 2016. Т. 293. № 02. С. 333–345.
21. *Lugosi G.* Comment on: ℓ_1-penalization for mixture regression models // TEST. 2010. V. 19. № 2. P. 259–263.
22. *Friedlander M.P., Tseng P.* Exact regularization of convex programs // SIAM J. Optim. 2008. V. 18. № 4. P. 1326–1350.
23. *Гасников А.В., Двуреченский П.Е., Дорн Ю.В., Максимов Ю.В.* Численные методы поиска равновесного распределения потоков в модели бэкмана и в модели стабильной динамики // Матем. моделирование. 2016. Т. 28. № 10. С. 40–64.
24. *Anikin A., Gasnikov A., Gornov A., Kamzolov D., Maximov Y., Nesterov Y.* Efficient numerical methods to solve sparse linear equations with application to pagerank // arXiv e-prints. 2015. arXiv:1508.07607.
25. *Cox B., Juditsky A., Nemirovski A.* Decomposition techniques for bilinear saddle point problems and variational inequalities with affine monotone operators // J. Optimiz. Theory Appl. 2016. V. 172. № 2. P. 402–435.
26. *Blumensath T., Davies M.E.* Gradient pursuits // IEEE T. Signal Process. 2008. V. 56. № 6. P. 2370–2382.
27. *Blumensath T., Davies M.E.* Stagewise weak gradient pursuits // IEEE T. Signal Process. 2009. V. 57. № 11. P. 4333–4346.
28. *Nesterov Y.* Introductory Lectures on Convex Optimization. Boston: Kluwer Academic Publishers, 2004.
29. *Nemirovski A.* Optimization II: Numerical methods for nonlinear continuous optimization. Lecture Notes. Israel Institute of Technology, 1999.
30. *Dereventsov A.V.* On the approximate weak chebyshev greedy algorithm in uniformly smooth banach spaces // J. Math. Anal. Appl. 2016. V. 436. № 1. P. 288–304.
31. *Nemirovski A., Onn S., Rothblum U.G.* Accuracy certificates for computational problems with convex structure // Math. Oper. Res. 2010. V. 35. № 1. P. 52–78.
32. *Гасников А.В.* Современные численные методы оптимизации, метод универсального градиентного спуска: учебное пособие // arXiv e-prints. 2017. arXiv:1711.00394.
33. *Nesterov Y.* Complexity bounds for primal-dual methods minimizing the model of objective function // Math. Program. 2017. V. 171. № 1–2. P. 311–330.