МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«КРЫМСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ имени В. И. Вернадского»

Ф. С. Стонякин

АДАПТИВНЫЕ АЛГОРИТМИЧЕСКИЕ МЕТОДЫ В НЕГЛАДКОЙ ОПТИМИЗАЦИИ

Монография

Симферополь — 2020

УДК 519.85

ББК 22.19

Печатается по решению Научно-технического совета Крымского федерального университета имени В. И. Вернадского, протокол № 2 от 3 июня 2020 г.

Адаптивные алгоритмические методы в негладкой оптимизации / доц. Ф. С. Стонякин. Симферополь: ПОЛИПРИНТ, 2020.

Аннотация

Ключевая цель монографии — изложение результатов автора, посвященных алгоритмическим методам для некоторых классов (преимущественно негладких) оптимизационных задач. Первый подход основан на введении в оптимизационную модель искусственных неточностей. Введены соответствующие аналоги известного понятия неточного оракула для вариационных неравенств, седловых задач и задач минимизации функционалов. Предложены методы градиентного типа с адаптивными процедурами выбора шагов и критериев остановки, применимые к выделенным классам задач. Второй подход связан с использованием субградиентных схем с переключениями для задач выпуклого программирования на достаточно продвинутом уровне с новыми адаптивными критериями остановки, что позволяет рассматривать некоторые классы задач с необязательно ограниченными субградиентами. Обсуждаются приложения полученных результатов к некоторым прикладным задачам (максимизация полезности компьютерной сети, проектирование механических конструкций и другие). Издание предназначено для студентов старших курсов с математической специализацией, аспирантов и научно-педагогических работников в области методов оптимизации.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Во многих прикладных задачах возникает необходимость подходящих алгоритмических методов для задач негладкой выпуклой оптимизации. Однако оценки эффективности таких процедур в случае большой размерности переменных весьма пессимистичны. Так, к примеру, ε-точное решение по функции задачи выпуклой негладкой оптимизации возможно достичь за $O(\varepsilon^{-2})$ обращений к подпрограмме нахождения (суб)градиента, и в общем случае такая оценка не улучшаема [36]. Для гладких задач оценки эффективности выше, что приводит к естественной идее для негладких задач обосновать возможность использования какого-нибудь приближения оптимизационной модели к гладкому случаю. Эту идею, в частности, реализуют так называемые *универсальные методы*, исследованию которых было положено начало в работе [152]. Универсальные градиентные методы основаны на построении для задач выпуклой оптимизации с гёльдеровым (суб)градиентом целевого функционала аналога стандартной квадратичной интерполяции с искусственно введённой погрешностью. Универсальность метода при этом понимается как возможность адаптивной настройки при работе метода на оптимальный в некотором смысле уровень гладкости задачи и величину, соответствующую константе Гёльдера L_ν (суб)градиента целевого функционала. Оказывается, что возможность такой настройки может позволить экспериментально для некоторых задач улучшить скорость сходимости по сравнению с оптимальными теоретическими оценками [152].

Искусственную неточность для негладких задач можно вводить поразному. При этом естественно возникает проблема описания влияния погрешностей задания целевого функционала и градиента на оценки скорости сходимости методов. Для градиентных методов известен подход к этой проблеме, основанный на недавно предложенной концепции неточного оракула [90, 91]. Известно, что для неускоренных градиентных методов в оценках скорости сходимости не происходит накопления величин, связанных с погрешностями. Однако при этом для оптимального при отсутствии погрешностей на классе гладких выпуклых задач быстрого градиентного метода в итоговой оценке скорости

сходимости величины погрешностей могут накапливаться. Концепция неточного оракула была обобщена в [22, 168], где были введены понятия (δ, L)-модели и (δ, L, μ)-модели целевой функции для задач оптимизации. Суть данных обобщений в замене $\langle \nabla f(x), y - x \rangle$ некоторой абстрактной выпуклой по первой переменной функцией $\psi(y, x)$, что позволяет рассматривать более широкий класс задач [168].

Различным методам градиентного типа посвящены все новые современные работы. В частности, недавно в [67] введены условия относительной гладкости оптимизируемого функционала, которые предполагают замену условия липшицевости градиента на ослабленный вариант:

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + LV(y, x),$$

где $V(y, x)$ — широко используемый в оптимизации аналог расстояния между точками x и y, который называют *дивергенцией Брегмана*. Обычно *дивергенция Брегмана* вводится на базе вспомогательной 1-сильно выпуклой функции d (порождает расстояния), дифференцируемой во всех точках выпуклого замкнутого множества Q:

$$V(y, x) = d(y) - d(x) - \langle \nabla d(x), y - x \rangle \quad \forall x, y \in Q,$$

где $\langle \cdot, \cdot \rangle$ — скалярное произведение в $\mathbb{R}^n$. В частности, для стандартной евклидовой нормы $\| \cdot \|_2$ и расстояния в $\mathbb{R}^n$ можно считать, что $V(y, x) = d(y - x) = \frac{1}{2}\|y - x\|_2^2$ для произвольных $x, y \in Q$. Однако часто возникает необходимость использовать и неевклидовы нормы. Более того, рассмотренное в [67, 124] условие относительной гладкости предполагает лишь выпуклость (но не сильную выпуклость) порождающей функции d. Как показано в [124], концепция относительной гладкости позволяет применить вариант градиентного метода для некоторых задач, которые ранее решались лишь с помощью методов внутренней точки. В частности, речь идет об известной задаче построения оптимального эллипсоида, покрывающего заданный набор точек. Эта задача, в частности, представляет интерес для статистики и анализа данных. Отметим в этой связи также предложенный недавно в [125] подход к задачам негладкой оптимизации, связанный с релаксацией условия Липшица, которая предполагает замену ограниченности нормы субградиента $\|\nabla f(x)\|_* \leqslant M_f$ так называемой *относительной липшицевостью*:

$$\|\nabla f(x)\|_* \leqslant \frac{M_f \sqrt{2V(y, x)}}{\|y - x\|} \quad \forall x, y \in Q,\ y \neq x.$$

При этом порождающая функция d не обязательно сильно выпукла. В работе [125] предложены детерминированный и стохастический алгоритмы зеркального спуска для задачи минимизации выпуклого относительно липшицева целевого функционала.

Как известно, погрешности при решении задач оптимизации возникают в силу разных причин [46]. Они могут быть естественно связанными с неточностью доступных данных, заменой бесконечномерной задачи конечномерным аналогом и т.д. Возможны и искусственные неточности, возникающие в ходе математического исследования рассматриваемых задач. Помимо указанной выше идеологии универсальных методов в этом плане можно отметить и неточности, связанные с техникой регуляризации задачи, а также со сглаживанием [149].

Поэтому естественно возникает проблема описания влияния погрешностей задания целевого функционала и градиента на оценки скорости сходимости методов. Для градиентных методов выпуклой оптимизации известен подход к этой проблеме, основанный на недавно предложенной концепции неточного оракула [90, 91]. Известно, что для обычного (неускоренного) градиентного метода в оценке скорости сходимости не происходит накопления величин, связанных с погрешностями. Однако для оптимальных при отсутствии погрешностей на классе гладких задач ускоренных методов (например, для быстрого градиентного метода) в итоговой оценке скорости сходимости величины погрешностей могут накапливаться. Известны подходы к этой проблеме для специальной концепции детерминированного шума (аппроксимативный градиент) [89], а также — для случайного аддитивного шума [16, 88, 102] при задании градиента.

Также весьма важной является задача построения методов для вариационных неравенств и седловых задач с адаптацией как уровню гладкости оператора, так и к величинам погрешностей. Отметим, что вариационные неравенства имеют широкий спектр приложений в физике, технике, экономике, а в последнее время и в машинном обучении. К примеру, отмечена целесообразность их применения в задачах описания генеративно-состязательных сетей [65]. В качестве аналога гладкой оптимизационной задачи можно рассматривать задачу нахождения решения вариационного неравенства с монотонным липшицевым оператором, а в качестве аналога негладкой оптимизационной задачи — вариационное неравенство с ограниченным оператором. Оптимальным с точ-

ки зрения нижних оценок количества обращений к оракулу оператора поля для указанных классов задач будет экстраградиентный метод [32], а также его более современный вариант — проксимальный зеркальный метод [135]. В работе [66] предложен адаптивный метод для вариационных неравенств со случайным шумом и поставлена задача разработки метода с адаптивной настройкой на уровень гладкости оператора с исследованием вопросов накопления в итоговой оценке качества найденного решения детерминированного аддитивного шума при задании оператора.

Настоящая работа посвящена проблеме построения адаптивных методов для задач выпуклой оптимизации, вариационных неравенств и седловых задач с адаптивной настройкой как на уровень гладкости целевого функционала (оператора), так и на величину детерминированного аддитивного шума. При этом ставится задача обосновать применимость разрабатываемых алгоритмических методов на классе задач со структурой, которая допускает их описание с точки зрения некоторой концепции абстрактной неточной модели оптимизируемой функции подобно [22, 168]

Цель работы — разработка оптимальных алгоритмических методов для задач выпуклой негладкой оптимизации, а также концепций неточных моделей для вариационных неравенств с монотонными операторами, выпукло-вогнутых седловых задач и задач оптимизации, которые позволили бы предложить эффективные алгоритмические методы с адаптацией оценок качества решения к уровню гладкости и величинам погрешностей.

Можно выделить такие **задачи** исследования, положенного в основу работы:

1. Распространение идеологии универсальных градиентных методов на существенно более широкие классы задач, а именно — на вариационные неравенства и седловые задачи. Разработка концепций абстрактных неточных оптимизационных моделей (аналоги (δ, L)-оракула и (δ, L)-модели функции) для вариационных неравенств и седловых задач, которые позволяют получить эффективные оценки скорости сходимости с учётом естественных и некоторых типов искусственных погрешностей для таких задач.

2. Исследование новых подходов к построению неточной оптимизационной модели целевого функционала для минимизационных задач с

раздельным учётом абсолютных погрешностей задания целевой функции и градиента (субградиента). Разработка для соответствующих классов задач алгоритмических методов, которые могли бы позволить более гибко учитывать степень негладкости задачи, а также возникающие естественные и искусственные неточности с сохранением приемлемых вычислительных гарантий.

3. Описание условий на целевой функционал оптимизационной задачи, которые позволяют гарантировать линейную скорость сходимости метода с точностью до величин, соответствующих погрешностям.

4. Исследование возможности распространения указанных подходов с сохранением оптимальных вычислительных гарантий на некоторые классы невыпуклых оптимизационных задач.

5. Разработка адаптивных алгоритмических схем зеркального спуска для задач негладкого выпуклого программирования с возможностью обоснования оптимальных вычислительных гарантий на классах задач с нелипшицевыми целевыми функционалами, возникающих в приложениях.

6. Тестирование предложенных алгоритмических методов для некоторых примеров оптимизационных задач.

Теоретические исследования по обоснованию необходимых результатов и оценок скорости сходимости разработанных методов были выполнены с использованием **методов** математического анализа, выпуклого и функционального анализа. Тестирование предложенных алгоритмов выполнено с использованием компьютерных программ в среде CPython 3.7. Все вычисления в работе в ходе экспериментов были произведены на компьютере с 3-ядерным процессором AMD Athlon II X3 450 с тактовой частотой 3,2 ГГц на каждое ядро. ОЗУ компьютера составляло 8 Гб.

Сформулируем наиболее важные полученные результаты:

1. Выделен новый класс задач вариационных неравенств и седловых задач, допускающих предложенный в диссертации вариант неточного оракула (абстрактной модели функции) с условиями типа относительной гладкости. На этом классе задач предложен адаптивный вариант проксимального зеркального метода, получена оценка его скорости сходимости и доказано, что в ней не накапливаются определяемые погрешностями величины. Тем самым, существенно расширены границы применимости методов экстраградиентного типа с сохранением опти-

мальных оценок сложности. Обоснована применимость данного метода к популярным для вопросов обработки изображений композитным седловым задачам.

2. Предложен единообразный подход к построению оптимизационной модели для вариационных неравенств с ν-гёльдеровыми монотонными операторами (при всех $\nu \in [0;1]$), который позволил разработать универсальный метод для вариационных неравенств т выпукло-вогнутых седловых задач соответствующего уровня гладкости. Получены оценки скорости сходимости этого метода, указывающие на оптимальность предложенного подхода как на классе вариационных неравенств с липшицевыми операторами ($\nu = 1$), так и на классе вариационных неравенств с ограниченными операторами ($\nu = 0$).

3. С использованием усреднения адаптивно подбираемых констант Липшица оператора на итерациях предложен адаптивный метод для вариационных неравенств с липшицевыми сильно монотонными операторами с гарантией линейной скорости сходимости.

4. Предложен новый подход к понятию абстрактной неточной оптимизационной модели (неточного оракула) для градиентных методов оптимизации, показана его применимость к достаточно широкому выделенному классу задач негладкой оптимизации. Особенность предложенного подхода — раздельный учёт влияния погрешностей разного типа путём введения несколькими параметров погрешностей. Предложен метод градиентного типа с адаптивной настройкой в оценке скорости сходимости некоторых из этих параметров и показано, что такая адаптивная настройка может повышать качество найденного решения. Доказано, что величины, связанные со всеми типами погрешностей, не накапливаются в итоговых оценках для неускоренных методов. Предложена дополнительная процедура для ускоренного метода, гарантирующая отсутствие накопления одного из типов погрешностей.

5. Введено новое понятие неточной модели (аналог неточного оракула) целевой функции с несколькими параметрами, соответствующими свойствам сильной выпуклости, гладкости, а также с раздельным учётом погрешностей задания целевого функционала и градиента. Обоснована близкая к линейной скорость сходимости предложенных градиентных методов с адаптивным выбором шага для задач минимизации функционалов, которые допускают существование указанного типа оптимизационной модели в произвольной запрошенной точке. Обоснова-

на применимость разработанной методики на некотором классе негладких задач оптимизации с вычислительными гарантиями, близкими к оптимальным. Разработан адаптивный градиентный метод и обоснована близкая к линейной скорость его сходимости в случае, если вместо сильной выпуклости целевого функционала относительно евклидовой нормы выполняется условие градиентного доминирования.

6. Комбинированием метода дихотомии с неточным решением одномерной двойственной задачи и методов градиентного типа для вспомогательных многомерных задач предложен подход к задачам выпуклого программирования с одним функциональными ограничением. Методика основана на введённом в работе адаптивном критерии остановки, который учитывает неточность решения вспомогательных задач. На базе полученной оценки скорости сходимости для аналога метода дихотомии на квадрате с учётом погрешности нахождении градиента предложен подход к задаче выпуклого программирования с двумя функциональными ограничениями. Показана оптимальность предложенной методики для достаточно гладких задач с сильно выпуклым целевым функционалом как в случае одного, так и двух функционалов ограничений.

7. Для зеркальных спусков с переключениями введены новые адаптивные критерии остановки, выполнение которых гарантирует достижение приемлемого качества решения задачи выпуклого программирования вне зависимости от уровня гладкости целевого функционалам. Обоснована оптимальность оценок скорости сходимости для некоторых классов негладких целевых функционалов, не удовлетворяющих условию Липшица. В частности доказано, что на классе задач с гёльдеровыми выпуклыми целевыми функционалами сохранится оценка сложности $O(\varepsilon^{-2})$, оптимальная даже на более узком классе липшицевых выпуклых целевых функционалов. Рассмотрено приложение к задаче оптимизации высоконагруженной компьютерной сети. Также на базе разработанных методов зеркального спуска предложены методы для задач минимизации с относительной точностью выпуклого однородного функционала с выпуклыми функционалами ограничений. Эти методы гарантируют достижение заданной относительной точности приближённого решения задачи за оптимальное число итераций при существенно более общих предположениях в сравнении с известными аналогичными результатами (0 — не обязательно внутренняя точка субдифференциала целевой функции в нулевой точке).

8. С помощью техники рестартов указанных в п. 7 адаптивных зеркальных спусков впервые разработаны методы для задач сильно выпуклой оптимизации с сильно выпуклым липшицевым функционалом ограничения. Впервые обоснована оптимальность рестартов зеркальных спусков с переключениями на некоторых классах негладких целевых функционалов, которые не удовлетворяю условию Липшица. В частности доказано, что на классе задач с гёльдеровыми сильно выпуклыми целевыми функционалами сохранится оценка сложности $O(\varepsilon^{-1})$, оптимальная даже на более узком классе липшицевых сильно выпуклых целевых функционалов.

9. Обоснована применимость некоторых из разработанных адаптивных зеркальных спусков с переключениями к задачам минимизации произвольного квазивыпуклого субдифференцируемого по Кларку целевого функционала с ненулевым субградиентом в любой точке и выпуклым функциональным ограничением. При этом показано, что сохраняются оценки сложности методов, которые оптимальны для более узкого класса задач выпуклого программирования с целевыми функционалами соответствующего уровня гладкости.

10. Предложен новый адаптивный метод зеркального спуска для задач онлайн-оптимизации в случае выпуклых (возможно, негладких) липшицевых целевых функционалов и нескольких выпуклых липшицевых функциональных ограничений. Обоснована оптимальность метода в терминах нижних оракульных оценок на рассмотренном классе задач.

Структура и основное содержание книги

Глава 1 содержит обзор литературных источников по теме, а также все необходимые вспомогательные результаты. В частности напоминается, что в основу обоснования оценок скорости сходимости методов градиентного типа может быть положена идея аппроксимации функции в исходной точке (текущем положении метода) мажорирующим ее параболоидом вращения. Так, для задачи минимизации выпуклого функционала $f : Q \to \mathbb{R}$ (если специально не оговорено иное, то всюду далее Q полагаем выпуклым замкнутым подмножеством $\mathbb{R}^n$) с липшицевым градиентом

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L \|x - y\| \quad \forall x, y \in Q$$

выполняются неравенства

$$f(x)+\langle\nabla f(x),y-x\rangle\leqslant f(y)\leqslant f(x)+\langle\nabla f(x),y-x\rangle+\frac{L}{2}\|y-x\|^2. \quad (0.1)$$

Неравенства (0.1) позволяют получить для обычного градиентного спуска оценку скорости сходимости

$$f(\widehat{x})-f^*\leqslant\frac{C_1}{N},$$

где $\widehat{x}$ — выход работы метода после N итераций, f^* — точное значение искомого минимума функции f, $C_1>0$ — некоторая постоянная.

Далее, приводятся необходимые вспомогательные сведения об известной концепции неточного оракула О. Деволдера–Ф. Глинера–Ю. Е. Нестерова для оптимизационных задач. Говорят, что функция f допускает неточный оракул $(f_\delta(x),g_\delta(x))\in\mathbb{R}\times E^*$, если выполняется некоторый аналог неравенства (0.1):

$$\begin{aligned}f_\delta(x)+\langle g_\delta(x),y-x\rangle\leqslant f(y)\leqslant\\ \leqslant f_\delta(x)+\langle g_\delta(x),y-x\rangle+\frac{L}{2}\|y-x\|^2+\delta\ \ \forall x,y\in Q.\end{aligned} \quad (0.2)$$

По сути, (0.2) означает, что $f_\delta(x)$ есть некоторое приближенное значение $f(x)$, а $g_\delta(x)$ — некоторый δ-субградиент f в точке x. Оказывается, что для неускоренного градиентного метода при выполнении условия (0.2) для градиентного метода (с заменой пары $(f,\nabla f)$ на (f_δ,g_δ)) верна такая оценка скорости сходимости:

$$f(\widehat{x})-f^*\leqslant\frac{C_1}{N}+2\delta, \quad (0.3)$$

т.е. соответствующие погрешностям величины не накапливаются. При этом для быстрого градиентного метода (с заменой пары $(f,\nabla f)$ на (f_δ,g_δ)) выполняется следующая оценка скорости сходимости:

$$f(\widehat{x})-f^*\leqslant\frac{C_2}{(N+1)^2}+N\delta,\ \text{для некоторой постоянной } C_2>0. \quad (0.4)$$

Сравнение оценок (0.3) и (0.4) указывает на неочевидность преимущества использования ускоренного метода при наличии погрешностей.

Идеология Деволдера–Глинера–Нестерова развита в недавней работе [22], где было предложено обобщение концепции (δ,L)-оракула —

(δ, L)-модель целевой функции. Суть подхода в том, что $\langle \nabla f(x), y - x \rangle$ в (0.2) заменяется на некоторую абстрактную выпуклую функцию $\psi(y, x)$.

Определение 0.0.1. Будем говорить, что функция f допускает (δ, L)-модель в точке $x \in Q$, и обозначать эту модель $(f_\delta(x), \psi(y, x))$, если для любого $y \in Q$ справедливо неравенство

$$0 \leqslant f(y) - f_\delta(x) - \psi(y, x) \leqslant \frac{L}{2}\|y - x\|^2 + \delta, \tag{0.5}$$

где $\psi(x, x) = 0 \ \ \forall x \in Q$ и $\psi(y, x)$ — выпуклая функция по y для всякого $x \in Q$.

Концепция из определения 0.0.1 позволяет обосновать сходимость градиентного метода выпуклой минимизации для достаточно широкого класса задач оптимизации [168]. По сути, указанный подход позволяет унифицировать подходы к различным на первый взгляд классам задач оптимизации с проработкой вопросов влияния погрешностей данных на гарантированное качество решения, которого можно достичь в ходе работы метода. В качестве естественных примеров задач, в которых можно использовать концепцию модели оптимизируемой функции, отметим популярные в анализе данных задачи композитной оптимизации вида $f(x) = g(x) + h(x) \to \min\limits_{x \in Q}$, где $g(x)$ — выпуклая и гладкая функция (с L-липшицевым градиентом), $h(x)$ — выпуклая функция простой структуры (не обязательно гладкая). В качестве конкретного примера можно рассмотреть проблему восстановления матрицы корреспонденций по замерам потоков на линках (ребрах) в большой компьютерной сети (Minimal Mutual Information Model), которая сводится к задаче композитной оптимизации: $f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \mu \sum\limits_{k=1}^{n} x_k \ln x_k \to \min\limits_{x \in S_n(1)}$, где $S_n(1)$ — единичный симплекс в n-мерном пространстве.

Глава 2 посвящена новой методике решения вариационных неравенств и седловых задач, которая позволяет учитывать погрешности задания оператора, а также применима к задачам различного уровня гладкости.

Во **введении к главе 2** напоминается постановка задачи решения вариационного неравенства, а также необходимые понятия и результаты. Для оператора $G : Q \to \mathbb{R}^n$, заданного на выпуклом компакте

$Q \subset \mathbb{R}^n$ под *вариационным неравенством* понимаем неравенство вида

$$\langle G(x_*),\ x_* - x\rangle \leqslant 0. \tag{0.6}$$

Отметим, что в (0.6) требуется найти $x_* \in Q$ (это x_* и называется (строгим) решением ВН), для которого

$$\max_{x\in Q}\langle G(x_*),\ x_* - x\rangle \leqslant 0.$$

Для монотонного оператора поля G можно рассматривать также задачу отыскания слабого решения ВН

$$\langle G(x),\ x_* - x\rangle \leqslant 0, \tag{0.7}$$

то есть нахождения $x_* \in Q$, такого, что (0.7) верно при всех $x \in Q$.

Предложен аналог концепции неточной модели целевой функции в оптимизации для вариационных неравенств и седловых задач. Для удобства будем рассматривать задачу нахождения решения $x_* \in Q$ абстрактной задачи равновесного программирования

$$\psi(x, x_*) \geqslant 0 \quad \forall x \in Q \tag{0.8}$$

для некоторого выпуклого компакта $Q \subset \mathbb{R}^n$, а также функционала $\psi : Q \times Q \to \mathbb{R}$. Если предположить абстрактную монотонность функционала ψ:

$$\psi(x, y) + \psi(y, x) \leqslant 0\ \ \forall x, y \in Q,$$

то всякое решение (0.8) будет также и решением двойственной задачи равновесного программирования

$$\psi(x_*, x) \leqslant 0 \quad \forall x \in Q. \tag{0.9}$$

В общем случае делается предположение о существовании решения x_* задачи (0.8). Приведем пару примеров задания ψ, для которых данное условие заведомо выполнено.

Пример 0.0.2. Если для некоторого оператора $G : Q \to \mathbb{R}^n$ положить

$$\psi(x, y) = \langle G(y), x - y\rangle\ \ \forall x, y \in Q,$$

то (0.8) и (0.9) будут равносильны соответственно стандартным сильному и слабому вариационному неравенству с оператором G.

Пример 0.0.3. Для некоторого оператора $G : Q \to \mathbb{R}^n$ и выпуклого функционала $h : Q \to \mathbb{R}^n$ простой структуры выбор

$$\psi(x, y) = \langle G(y), x - y\rangle + h(x) - h(y)$$

приводит к *смешанному вариационному неравенству*

$$\langle G(y), y - x\rangle + h(y) - h(x) \leqslant 0,$$

которое в случае монотонности оператора G влечет

$$\langle G(x), y - x\rangle + h(y) - h(x) \leqslant 0.$$

В разделе 2.1 вводится концепция (δ, L)-модели для указанного выше класса задач и предлагается аналог проксимального зеркального метода А.С. Немировского с адаптивным выбором шага.

Определение 0.0.4. Будем говорить, что функционал ψ допускает (δ, L)-модель $\psi_\delta(x, y)$ при некоторых фиксированных $L > 0$ и $\delta > 0$ в произвольной точке y относительно дивергенции Брегмана $V(y, x)$, если для всяких $x, y, z \in Q$ верны:

(i) $\psi(x, y) \leqslant \psi_\delta(x, y) + \delta$;

(ii) $\psi_\delta(x, y)$ выпуклый по первой переменной;

(iii) $\psi_\delta(x, x) = 0$;

(iv) (*абстрактная δ-монотонность*)

$$\psi_\delta(x, y) + \psi_\delta(y, x) \leqslant \delta;$$

(v) (*обобщённая относительная гладкость*)

$$\psi_\delta(x, y) \leqslant \psi_\delta(x, z) + \psi_\delta(z, y) + LV(x, z) + LV(z, y) + \delta. \qquad (0.10)$$

Естественно возникает идея обобщить этот метод на абстрактные задачи (0.8) и (0.9) в предположениях их разрешимости, а также (i)–(iv). При этом будем учитывать погрешность δ в (0.10), а также погрешность $\widetilde{\delta}$ решения вспомогательных задач на итерациях согласно одному

из достаточно известных в алгоритмической оптимизации подходов (подробное обсуждение имеется, к примеру, в разделе 3 из [16]):

$$x := \arg\min_{y\in Q}^{\widetilde{\delta}} \varphi(y), \text{ если } \langle \nabla\varphi(x), x-y\rangle \leqslant \widetilde{\delta} \quad \forall y \in Q. \tag{0.11}$$

Опишем $(N+1)$-ую итерацию предложенного метода ($N = 0, 1, 2, \ldots$), выбрав начальное приближение $x^0 = \arg\min\limits_{x\in Q} d(x)$, зафиксировав точность $\varepsilon > 0$, а также некоторую константу $L_0 \leqslant 2L$.

Алгоритм 1 Адаптивный метод для концепции (δ, L)-модели для ВН.

---

1. $N := N + 1$; $L_{N+1} := \frac{L_N}{2}$.

2. Вычисляем:
   $y^{N+1} := \arg\min\limits_{x\in Q}^{\widetilde{\delta}} \left\{\psi_\delta(x, x^N) + L_{N+1}V(x, x^N)\right\}$,
   $x^{N+1} := \arg\min\limits_{x\in Q}^{\widetilde{\delta}} \left\{\psi_\delta(x, y^{N+1}) + L_{N+1}V(x, x^N)\right\}$
   до тех пор, пока не будет выполнено:

$$\begin{aligned}\psi_\delta(x^{N+1}, x^N) \leqslant \psi_\delta(y^{N+1}, x^N) + \psi_\delta(x^{N+1}, y^{N+1}) + \\ + L_{N+1}V(y^{N+1}, x^N) + L_{N+1}V(x^{N+1}, y^{N+1}) + \delta.\end{aligned} \tag{0.12}$$

3. **Если** (0.12) не выполнено, **то** $L_{N+1} := 2L_{N+1}$ и повторяем п. 2.

4. **Иначе** переход к п. 1.

5. Критерий остановки метода:

$$S_N := \sum_{k=0}^{N-1} \frac{1}{L_{k+1}} \geqslant \frac{\max\limits_{x\in Q} V(x, x^0)}{\varepsilon}.$$

Справедлива следующая

Теорема 0.0.5. *После остановки алгоритма 1 для всякого $x \in Q$ будет выполнено неравенство:*

$$-\frac{1}{S_N}\sum_{k=0}^{N-1}\frac{\psi_\delta(x, y^{k+1})}{L_{k+1}} \leqslant \frac{V(x, x^0)}{S_N} + 2\widetilde{\delta} + \delta \leqslant \varepsilon + 2\widetilde{\delta} + \delta, \tag{0.13}$$

а также

$$\psi(\widetilde{y}, x) \leqslant \frac{V(x, x^0)}{S_N} + 2\widetilde{\delta} + 3\delta \leqslant \varepsilon + 2\widetilde{\delta} + 3\delta \tag{0.14}$$

при

$$\widetilde{y} := \frac{1}{S_N} \sum_{k=0}^{N-1} \frac{y^{k+1}}{L_{k+1}}. \tag{0.15}$$

Замечание 0.0.6. Ввиду (0.10) и выбора $L_0 \leqslant 2L$ гарантированно будет верно $L_{k+1} \leqslant 2L \ \ \forall k = \overline{0, N-1}$. Поэтому $S_N \geqslant \frac{N}{2L}$ и оценки (0.13)–(0.14) означают, что для всякого $x \in Q$ будут верны неравенства:

$$\psi(\widetilde{y}, x) \leqslant \varepsilon + 2\widetilde{\delta} + 3\delta \tag{0.16}$$

после выполнения не более, чем $O\left(\varepsilon^{-1}\right)$ итераций предлагаемого метода. При этом нетрудно проверить, количество решений вспомогательных задач в п.2 алгоритма на N итерациях метода не превышает $2N + \log_2 \frac{L}{L_0}$, т.е. стоимость итерации в среднем будет сопоставимой со стоимостью итерации классического экстраградиентного метода, предполагающей решение двух вспомогательных задач на каждой итерации. Отметим, что оценка с точностью до числового множителя оптимальна для вариационных неравенств с монотонным липшицевым оператором [134].

Замечание 0.0.7. Для обычных слабых вариационных неравенств (0.7) неравенство (0.16) можно заменить на

$$\max_{x \in Q} \langle G(x), \widetilde{y} - x \rangle \leqslant \varepsilon + 2\widetilde{\delta} + 3\delta. \tag{0.17}$$

Отметим, что именно (0.17) обычно используют как критерий качества решения вариационного неравенства.

В разделе 2.2 рассмотрен случай гёльдерова оператора поля вариационного неравенства G

$$||G(x) - G(y)||_* \leqslant L_\nu ||x - y||^\nu \quad \forall x, y \in Q \tag{0.18}$$

для произвольного $\nu \in [0; 1]$, причем $L_0 < +\infty$ (другие константы L_ν ($\nu \neq 0$) могут быть бесконечными) Можно показать, что при условии (0.18)

$$\langle G(z) - G(y), z - x \rangle \leqslant \frac{L}{2}||z - x||^2 + \frac{L}{2}||z - y||^2 + \frac{\varepsilon}{2} \tag{0.19}$$

для константы

$$L = L_\nu^{\frac{2}{1+\nu}} (2\varepsilon)^{\frac{\nu-1}{1+\nu}},$$

которая зависит от введённой искусственной неточности ε. На базе интерполяции (0.19) алгоритм 1 с заменой $V(y,x)$ на $\frac{1}{2}\|y-x\|^2$ сводится к *универсальному* методу для вариационных неравенств, который предполагает адаптивную настройку на уровень гладкости оператора G. Доказано, что в таком случае при $L_\nu < +\infty$ разработанный метод позволяет получить приближённое решение задач (0.6) – (0.7) с точностью ε с оценкой достаточного количества итераций для достижения приемлемого качества решения

$$O\left(\varepsilon^{-\frac{2}{1+\nu}}\right), \tag{0.20}$$

которая оптимальна с точностью до постоянного множителя при $\nu = 0$ и $\nu = 1$. При этом адаптивность метода на практике может приводить к повышению скорости работы метода по сравнению с (0.20), что продемонстрировано на примере экспериментов.

В разделе 2.3 введён аналог понятия (δ, L)-модели для более узкого по сравнению с предыдущим разделом класса седловых задач и выписана оценка скорости сходимости алгоритма 1. Обоснована применимость этой концепции к композитным седловым задачам, которые возникают в задачах обработки изображений [84].

В разделе 2.4 приведены некоторые численные эксперименты, демонстрирующие преимущества использования именно адаптивных (универсальных) методов для рассмотренных в предыдущих разделах главы 2 классов задач. В частности, рассматриваются расчёты лагранжевых седловых задач для негладкой геометрической задачи Ферма–Торричелли–Штейнера с функциональными ограничениями, а также матричные игры.

В **разделе 2.5** с использованием методики усреднений адаптивно подбираемых констант Липшица оператора поля получена оценка скорости сходимости для предложенного автором адаптивного аналога метода [38, 143] для вариационных неравенств с сильно монотонным липшицевым оператором. Этот метод представляет собой комбинацию двойственного экстраполяционного метода [142] и методов оценочных функций.

Будем рассматривать задачу нахождения решения $x_* = x_*(Q)$ вари-

ационного неравенства

$$\langle G(x_*), x_* - y\rangle \leqslant 0 \quad \forall y \in Q, \tag{0.21}$$

где $G : Q \to \mathbb{R}^n$ — сильно монотонный оператор с параметром $\mu > 0$:

$$\langle G(x) - G(y), x - y\rangle \geqslant \mu\|x - y\|^2 \quad \forall x, y \in Q, \tag{0.22}$$

Q — выпуклое замкнутое подмножество $\mathbb{R}^n$, $\langle\cdot,\cdot\rangle$ — скалярное произведение в $\mathbb{R}^n$, $\|x\| = \langle Bx, x\rangle^{1/2}$ есть некоторая евклидова норма в $\mathbb{R}^n$, где $B : \mathbb{R}^n \to \mathbb{R}^n$ — фиксированный самосопряжённый оператор $B = B^T > 0$. Будем полагать, что оператор G удовлетворяет условию Липшица:

$$\|G(x) - G(y)\|_* \leqslant L\|x - y\| \ \forall x, y \in Q \tag{0.23}$$

для некоторой константы $L > 0$, $\|s\|_* = \langle s, B^{-1}s\rangle^{1/2}$.

Напомним некоторые вспомогательные оценки, понятия и результаты из п. 3.2 диссертации Ю. Е. Нестерова [38]. Отметим, что сильная монотонность G означает, что для решения x_* верны оценки при произвольном $y \in Q$:

$$\langle G(y), x_* - y\rangle + \frac{\mu}{2}\|y - x_*\|^2 \leqslant \langle G(x_*), x_* - y\rangle - \frac{\mu}{2}\|y - x_*\|^2 \leqslant 0. \tag{0.24}$$

Неравенства (5.70) приводят к идее рассматривать следующую меру близости для оценки качества найденного приближённого решения x ВН (5.71):

$$\rho(x) = \sup_{y \in Q}\left\{\langle G(y), x - y\rangle + \frac{\mu}{2}\|y - x\|^2\right\}. \tag{0.25}$$

Отметим основные свойства ρ из (0.25).

Теорема 0.0.8. *Функция ρ из* (0.25) *определена и сильно выпукла на Q с параметром μ. Более того, для всякого $x \in Q$ $\rho(x) \geqslant 0$ и $\rho(x) = 0 \Leftrightarrow x = x_*$.*

Пусть в ходе работы некоторого алгоритма образовалась последовательность $\{y_i\}_{i=0}^N \subset Q$ и $\{\lambda_i\}_{i=0}^N$ — некоторый набор положительных чисел. Тогда обозначим $S_N = \sum\limits_{i=0}^N \lambda_i$ и $\widetilde{y}_N := \frac{1}{S_N}\sum\limits_{i=0}^N \lambda_i y_i$ — усредненный выход работы алгоритма.

Неравенства (5.70) приводят к идее ввести следующую функцию зазора для оценки качества найденного решения:

$$\widehat{\Delta}_N := \max_{x\in Q}\left\{\sum_{i=0}^{N}\lambda_i\left[\langle G(y_i), y_i - x\rangle - \frac{\mu}{2}\|x-y_i\|^2\right]\right\}. \tag{0.26}$$

Лемма 0.0.9. *Справедливо неравенство:* $\rho(\widetilde{y}_N) \leqslant \frac{\widehat{\Delta}_N}{S_N}$.

Обозначим

$$\varphi_y^\beta(x) := \langle G(y), y - x\rangle - \frac{\beta}{2}\|x-y\|^2, \quad \Phi_k(x) := \sum_{i=0}^{k}\lambda_i\varphi_{y_i}^\mu(x)$$

для произвольного параметра $\beta > 0$, $k = 0, 1, 2, \ldots$, а также $x, y \in Q$. Ясно, что функция φ_y^β сильно вогнута с параметром β, а $\Phi_k(x)$ сильно вогнута с параметром μS_k. Заметим, что при этом $(k = 0, 1, 2, \ldots, N)$

$$\widehat{\Delta}_k = \max_{x\in Q}\Phi_k(x).$$

Напомним метод [38] для ВН с липшицевым сильно монотонным оператором. Опишем $(k+1)$-ю итерацию этого метода $(k = 0, 1, 2, \ldots)$.

Алгоритм 3 Метод для ВН с сильно монотонным оператором

$x_k := \arg\max\limits_{x\in Q}\Phi_k(x)$, $y_{k+1} := \arg\max\limits_{x\in Q}\varphi_{x_k}^L(x)$, $\lambda_{k+1} := \frac{\mu}{L}S_k$.

Выход: $\widetilde{y}_{k+1} := \frac{1}{S_{k+1}}\sum\limits_{i=0}^{k+1}\lambda_i y_i$.

Предложен адаптивный аналог алгоритма 3 для задач (5.71) – (0.22). Положим изначально $\lambda_0 := 1$, y_0 — некоторое начальное приближение искомого решения и выберем некоторое $0 < \beta_0 \leqslant 2L$, где L — константа Липшица для оператора G из (1.17).

Замечание 0.0.10. Ввиду сильной монотонности оператора G для всяких $x \neq y$ из множества Q верно $G(x) \neq G(y)$. Поэтому выполнения условия $\beta_0 \leqslant 2L$ можно добиться, выбрав

$$\beta_0 := \frac{\|G(x) - G(y)\|_*}{\|x-y\|}$$

для некоторых фиксированных различных x и y из Q.

Алгоритм 4 Адаптивный метод для ВН с сильно монотонным оператором

$x_k := \arg\max\limits_{x \in Q} \Phi_k(x)$, $\beta_{k+1} := \frac{\beta_k}{2}$, $y_{k+1} := \arg\max\limits_{x \in Q} \varphi_{x_k}^{\beta_{k+1}}(x)$.
Если верно

$$\|G(y_{k+1}) - G(x_k)\|_* \leqslant \sqrt{\beta_{k+1}(\beta_{k+1} + \mu)} \cdot \|y_{k+1} - x_k\|, \tag{0.27}$$

то вычисляем $\lambda_{k+1} := \frac{\mu}{\beta_{k+1}} S_k$, увеличиваем k на 1 и переходим к следующей итерации (п. 1).
Иначе $\beta_{k+1} := 2 \cdot \beta_{k+1}$ и переходим к п. 2.
Выход: $\widetilde{y}_{k+1} := \frac{1}{S_{k+1}} \sum\limits_{i=0}^{k+1} \lambda_i y_i$.

Опишем $(k+1)$-ю итерацию предлагаемого метода $(k = 0, 1, 2, \ldots)$.

Замечание 0.0.11. При $\beta_{k+1} \geqslant L$ критерий выхода из итерации (0.27) заведомо выполнен, т.к. $\sqrt{L(L+\mu)} > L$ при всяком $\mu > 0$.

За счёт повторения вычислений в п. 2 сложность работы предлагаемого алгоритма 4 по сравнению с алгоритмом 3 может увеличится не более, чем в 2 раза с точностью до постоянного слагаемого, зависящего от β_0 и L. Это означает, что трудоёмкость предлагаемого метода вполне сопоставима с трудоёмкостью исходного алгоритма 3. Однако при этом не требуется знания никакой константы $\widehat{L} \geqslant L$. Преимуществом также является возможное существенное увеличение скорости сходимости метода в конкретных задачах, что проиллюстрировано экспериментами. Справедлива следующая

Теорема 0.0.12. *При выполнении алгоритма 4 для величин* $\widehat{\Delta}_k$ *из* (5.69) *верно неравенство* $\widehat{\Delta}_{k+1} \leqslant \widehat{\Delta}_k$ *для всякого целого неотрицательного* k.

Следствие 0.0.13. *При выполнении алгоритма 4 верно неравенство* $f(\widetilde{y}_k) \leqslant \widehat{\Delta}_0 \exp\left(-\frac{k\mu}{\mu+\widehat{\beta}}\right)$ *для всякого натурального* k, *где* $\widehat{\beta}$ *определяется следующим образом:*

$$1 - \frac{\mu}{\mu + \widehat{\beta}} = \sqrt[k]{\left(1 - \frac{\mu}{\mu + \beta_1}\right)\left(1 - \frac{\mu}{\mu + \beta_2}\right) \ldots \left(1 - \frac{\mu}{\mu + \beta_k}\right)}.$$

Из теорем 5.7.44 и 0.0.12, а также следствия 0.0.13 вытекает следующий результат.

Теорема 0.0.14. *Пусть оператор G липшицев с константой $L > 0$ и сильно монотонен с параметром $\mu > 0$. Тогда при выполнении алгоритма 17 для $\gamma = \frac{L}{\mu}$ и всякого натурального k верны оценки:*

$$\frac{\mu}{2}\|\widetilde{y}_k - x_*\|^2 \leqslant \rho(\widetilde{y}_k) \leqslant$$

$$\leqslant \left[\rho(y_0) + \frac{\mu(\gamma^2-1)}{2}\|y_0 - x_*\|^2\right] \exp\left(-\frac{k}{1+\frac{\widehat{\beta}}{\mu}}\right) \leqslant$$

$$\leqslant \rho(y_0)\cdot\gamma^2\cdot\exp\left(-\frac{k}{1+\frac{\widehat{\beta}}{\mu}}\right).$$

Отметим, что полученные оценки могут оказаться лучше оценок для неадаптивного варианта метода, поскольку $\frac{\widehat{\beta}}{\mu}$ может оказаться меньше γ. Это наглядно продемонстрировано на примере численных экспериментов.

Глава 3 посвящена новым результатам в области методов градиентного типа с адаптивной настройкой на величины погрешностей данных, которые могут быть как естественными (приближённые значения целевой функции или градиента), так и искусственными (связаны с более низким уровнем гладкости целевого функционала). Отличительная особенность подхода — наличие не одного, а двух соответствующих возможным погрешностям параметров (в частности, для гладких задач погрешности задания целевого функционала и его градиента или субградиента). При этом удалось в некотором смысле обосновывать возможность избежать накопления значений одного из этих параметров погрешностей (задания градиента в гладком случае) в теоретических оценках скорости сходимости метода. Подчеркнём, что основная общая идея всех обсуждаемых в данной главе подходов — возможность адаптивной настройки не только на параметр гладкости, но и на величины параметров неточностей оптимизационной модели. Это позволяет уменьшить влияние погрешностей на итоговые оценки скорости сходимости, что проиллюстрировано результатами некоторых вычислительных экспериментов.

Во **введении** приводится мотивировка для разрабатываемого подхода к аналогу концепции (δ, L)-модели целевой функции с нескольки-

ми параметрами, которые описывают разные типы неточностей. Так, например для стандартной модели $\psi(y,x) = \langle \nabla f(x), y-x\rangle$ описывается ситуация некоторой модификации условий (0.5) с учетом отдельно погрешности задания f и ∇f. Если положить, что $\forall x \in Q$

$$\left\|\nabla f(x) - \widetilde{\nabla} f(x)\right\| \leqslant \Delta, \;\; \Delta > 0$$

для некоторого доступного приближенного значения $\widetilde{\nabla} f(x)$ градиента ∇f, то будет верно неравенств $\left|\left\langle \nabla f(x) - \widetilde{\nabla} f(x), y-x\right\rangle\right| \leqslant \Delta \|y-x\|$, то есть для всяких $x, y \in Q$

$$f(y) \leqslant f(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle + \frac{L}{2}\|y-x\|^2 + \Delta\|y-x\|,$$

а также

$$f(y) \geqslant f(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle - \Delta\|y-x\|.$$

Если кроме этого учесть неравенство $f_\delta(x) \leqslant f(x) \leqslant f_\delta(x) + \delta$ при $\delta > 0$, то получим следующий аналог (0.2):

$$f_\delta(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle - \Delta\|y-x\| \leqslant f(y) \leqslant$$
$$\leqslant f_\delta(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle + \frac{L}{2}\|y-x\|^2 + \Delta\|y-x\| + \delta \;\; \forall x, y \in Q$$

или

$$f(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle - \delta - \Delta\|y-x\| \leqslant f(y) \leqslant$$
$$\leqslant f(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle + \frac{L}{2}\|y-x\|^2 + \Delta\|y-x\| + \delta \;\; \forall x, y \in Q. \tag{0.28}$$

Можно ввести (в модельной общности подобно определению 0.0.1) следующий аналог неравенства (0.28) с параметрами $\delta, \gamma, \Delta \geqslant 0$:

$$f(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle - \delta - \gamma\|y-x\| \leqslant f(y) \leqslant$$
$$\leqslant f(x) + \left\langle \widetilde{\nabla} f(x), y-x\right\rangle + \frac{L}{2}\|y-x\|^2 + \Delta\|y-x\| + \delta \;\; \forall x, y \in Q. \tag{0.29}$$

Смысл такого обобщения заключается в том, что возможны различные значения параметров γ и Δ в (0.29) и, как показано в данной работе,

влияние величины Δ на итоговое качество решения может быть уменьшено. Отметим, что подробно разобрано несколько примеров негладких задач, когда $\delta = \gamma = 0$ при $\Delta > 0$. Если положить $\gamma = 0$, то $\widetilde{\nabla} f(x)$ — δ-субградиент f в точке x и параметр $\Delta > 0$ может указывать в этом случае на скачки $\widetilde{\nabla} f(x)$ в точках негладкости f. Вполне естественно рассматривать Δ как искусственную неточность для негладкой задачи. В таком случае Δ в неравенстве

$$\begin{aligned} f(x) + \left\langle \widetilde{\nabla} f(x), y - x \right\rangle \leqslant f(y) \leqslant \\ \leqslant f(x) + \left\langle \widetilde{\nabla} f(x), y - x \right\rangle + \frac{L}{2} \|y - x\|^2 + \Delta \|y - x\| \end{aligned} \tag{0.30}$$

вполне можно рассматривать как оценку скачков субдифференциалов f вдоль всевозможных векторных отрезков $[x; y]$. Несложно понять, что если f в (0.30) удовлетворяет условию Липшица с константой $M > 0$, то $\Delta \leqslant 2M$. Отметим, что возможна ситуация, когда Δ существенно меньше M. Похожие на (0.30) условия рассматривались в [120] для случая, когда f представим в виде суммы гладкого и негладкого целевого функционала. В настоящей работе рассмотрен более широкий класс целевых функционалов, которые не обязательно представимы в виде суммы гладкого и негладкого слагаемых.

В **разделе 3.1** предложено следующее максимально общее понятие неточной модели целевой функции, которая могла бы описать все указанные ситуации.

Определение 0.0.15. Будем говорить, что f допускает (δ, Δ, L)-модель в точке $x \in Q$, если для некоторой выпуклой по первой переменной функции $\psi(y, x)$ такой, что $\psi(x, x) = 0$, будет верно неравенство

$$\begin{aligned} \leqslant f_\delta(x) + \psi(y, x) \leqslant f(y) \leqslant \\ \leqslant f_\delta(x) + \psi(y, x) + \delta + \Delta \|y - x\| + LV(y, x) \end{aligned} \tag{0.31}$$

для произвольных $x, y \in Q$.

Покажем пример, поясняющий смысл использования модельной общности в предыдущем определении.

Пример 0.0.16. Отметим задачу выпуклой композитной оптимизации $f(x) = g(x) + h(x) \to \min$, где g — гладкая выпуклая функция, а h — выпуклая не обязательно гладкая функция простой структуры

(операция проектирования на любое множество уровня h не сильно затратна). Если при этом для градиента ∇g задано его приближение $\widetilde{\nabla} g$: $\left\|\widetilde{\nabla} g(x) - \nabla g(x)\right\| \leqslant \Delta$, причем

$$g(y) \geqslant g(x) + \langle \widetilde{\nabla} g(x), y - x\rangle - \gamma\|y - x\| - \delta,$$

то можно положить $\psi(y, x) = \langle \widetilde{\nabla} g(x), y - x\rangle + h(y) - h(x)$ и будет верно (0.31).

Предложен следующий метод, допускающий адаптацию оценки качества найденного решения к некоторым из параметров неточной модели.

Алгоритм 6 Адаптивный градиентный метод для функций, допускающих (δ, Δ, L)-модель в запрошенной точке ($\gamma = 0$).

Require: $x^0 \in Q$ — начальная точка, $V(x_*, x^0) \leqslant R^2$, параметры: $L_0 > 0,\ \Delta_0 > 0,\ \delta_0 > 0$: $L_0 \leqslant 2L,\ \Delta_0 \leqslant 2\Delta,\ \delta_0 \leqslant 2\delta$.

1: $L_{k+1} := {}^{L_k}/_2$, $\Delta_{k+1} := {}^{\Delta_k}/_2$, $\delta_{k+1} := {}^{\delta_k}/_2$.

2: $x^{k+1} := \arg\min\limits_{x \in Q}\{\psi(x, x^k) + LV(x, x^k)\}$.

3: **if** $f_\delta(x^{k+1}) \leqslant f_\delta(x^k) + \psi(x^{k+1}, x^k) + L_{k+1}V(x^{k+1}, x^k) + \Delta_{k+1}\left\|x^{k+1} - x^k\right\| + \delta_{k+1}$ **then**

4: $k := k + 1$ и выполнение п. 1.

5: **else**

6: $L_{k+1} := 2 \cdot L_{k+1}$; $\Delta_{k+1} := 2 \cdot \Delta_{k+1}$; $\delta_{k+1} := 2 \cdot \delta_{k+1}$ и выполнение п. 2.

7: **end if**

Ensure: $\widehat{x} := \frac{1}{S_N}\sum\limits_{k=0}^{N-1} \frac{x^{k+1}}{L_{k+1}}$, $S_N := \sum\limits_{k=0}^{N-1} \frac{1}{L_{k+1}}$.

Справедливо следующее утверждение.

Теорема 0.0.17. *Пусть $f : Q \to \mathbb{R}$ — выпуклая функция и $V(x_*, x^0) \leqslant R^2$, где x^0 — начальное приближение, а x_* — точное решение, ближайшее к x^0 с точки зрения дивергенции Брегмана. Тогда после N итераций для выхода $\widehat{x}$ алгоритма 6 будет верно неравенство*

$$f(\widehat{x}) - f(x_*) \leqslant \frac{R^2}{S_N} + \frac{1}{S_N}\sum_{k=0}^{N-1} \frac{\delta_{k+1} + \Delta_{k+1}\left\|x^{k+1} - x^k\right\|}{L_{k+1}} + \delta. \qquad (0.32)$$

Отметим, что вспомогательная задача п. 2 листинга алгоритма 6 решается не более

$$2N + \max\left\{\log_2 \frac{2L}{L_0}, \log_2 \frac{2\delta}{\delta_0}, \log_2 \frac{2\Delta}{\Delta_0}\right\} \tag{0.33}$$

раз.

Замечание 0.0.18. Как видим, не все параметры рассмотренной концепции неточной модели допускают адаптивную настройку. Поэтому по сути предложено лишь частичное решение проблемы разработки алгоритмического метода с адаптивной настройкой в оценке скорости сходимости параметров, соответствующих разным типам неточностей.

Замечание 0.0.19. Оценка (0.33) показывает, что в среднем трудоемкость итерации предложенного адаптивного алгоритма превышает трудоемкость неадаптивного метода не более, чем в постоянное число раз. Отметим также, что при $k = 0, 1, 2, \ldots$ $L_{k+1} \leqslant 2CL$, $C = \max\left\{1, \frac{2\delta}{\delta_0}, \frac{2\Delta}{\Delta_0}\right\}$. Поэтому $S_N \leqslant \frac{N}{2CL}$, что указывает на скорость сходимости метода $O(\varepsilon^{-1})$, но при наличии в оценке (0.32) слагаемых, определяемой параметрами δ и Δ (при этом ввиду адаптивности метода δ_k и Δ_k могут быть меньше δ и Δ соответственно).

Замечание 0.0.20. Отметим, что ввиду адаптивности алгоритма 6, полученная в теореме 0.0.17 оценка скорости сходимости может быть применена даже в случаях $L = +\infty$, $\Delta = +\infty$ или $\delta = +\infty$. Если не происходит зацикливание и каждый раз выполняется критерий выхода из итерации, то алгоритм 6 применим и в указанных ситуациях.

Далее, в **разделе 3.2** предложен вариант быстрого градиентного метода для задач выпуклой минимизации с адаптивным выбором шага и адаптивной настройкой на величины параметров (δ, Δ, L)-модели и получена оценка качества найденного решения. При этом возможна адаптивная настройка некоторых из параметров модели, соответствующих погрешностям задания целевой функции и градиента. Известные подходы к проблеме накопления ошибки градиента для ускоренного метода основаны на нетривиальной концепции аппроксимативного градиента [89], или на понимании погрешности как случайной величины [16,88,102]. Подход настоящей работы позволяет получить результат о скорости сходимости варианта ускоренного метода в модельной общности.

В **разделе 3.3** рассмотрен специальный класс задач выпуклой негладкой оптимизации, к которым применима концепция определения 0.0.15 ($\delta = 0$, $\Delta > 0$). Показано, что для этой ситуации возможно модифицировать неускоренный метод так, чтобы гарантированно достигалось ε-точное решение задачи минимизации f за

$$O\left(\frac{L}{\varepsilon}\right) + O\left(\frac{\Delta^2}{\varepsilon^2}\right) \tag{0.34}$$

обращений к (суб)градиенту целевого функционала. Для ускоренного метода аналогичной модификацией можно добиться того, чтобы гарантированно достигалось ε-точное решение задачи минимизации f за

$$O\left(\sqrt{\frac{L}{\varepsilon}}\right) + O\left(\frac{\Delta^2}{\varepsilon^2}\right) \tag{0.35}$$

обращений к (суб)градиенту целевого функционала. По сути получен некоторый аналог результата диссертации О. Деволдера (п. 4.7.2) [91]. Однако важным и новым является то, уже необязательно оптимизируемый функционал имеет вид суммы гладкого и негладкого слагаемого и рассмотрена модельная общность. Отметим, что в случае целевого функционала, равного сумме гладкого и M-липшицева функционалов, параметр Δ может быть существенно меньше M [55]. Также впервые предлагается метод с адаптивным выбором шага, имеющий гарантированные оценки скорости сходимости, близкие к (0.34) и (0.35). Отметим, что при этом для адаптивного ускоренного метода оценка может увеличиться на некоторый логарифмический множитель вида $O(log_2(\varepsilon^{-3}))$, для адаптивного неускоренного — на множитель вида $O(log_2(\varepsilon^{-1}))$.

Покажем, как возможно получить указанный результат для неускоренного метода в предположении, что целевой функционал допускает (δ, Δ, L)-модель ψ. Заметим, что при этом требуется 1-сильная выпуклость прокс-функции в определении дивергенции Брегмана.

Пусть на (k+1)-й итерации алгоритма 6 ($k = 0, 1, \ldots, N-1$) верно неравенство $L \leqslant L_{k+1} \leqslant 2L$ (как показано в п. 2 доказательства теоремы 0.0.17, этого можно всегда добиться выполнением не более чем постоянного числа операций п. 2 листинга алгоритма 6). Для каждой итерации алгоритма 6 ($k = 0, 1, \ldots, N-1$) предложим такую процедуру:

Повторяем операции п. 2 p раз, увеличивая L_{k+1} в два раза при неизменной $\Delta_{k+1} \leqslant 2\Delta$.	(0.36)

Процедуру (0.36) остановим в случае выполнения одного из неравенств:

$$\Delta_{k+1} \left\| x^{k+1} - x^k \right\| \leqslant \frac{\varepsilon}{2},$$

или

$$f(x^{k+1}) \leqslant f(x^k) + \psi(x^{k+1}, x^k) + 2^{p-1} L \left\| x^{k+1} - x^k \right\|^2.$$

Отметим, что здесь мы полагаем f точно заданной, то есть $f_\delta = f$ ($\delta = 0$) и $\psi(y, x) = \langle \nabla f(x), y - x \rangle$, где ∇f — некоторый субградиент f. Получена верхняя оценка на p и доказана следующая

Теорема 0.0.21. *Для выхода $\widehat{x}$ модифицированного алгоритма 6 с учетом дополнительной процедуры* (0.36) *неравенство $f(\widehat{x}) - f^* \leqslant \varepsilon$ будет гарантированно выполнено не более, чем после*

$$\left\lceil \frac{4LR^2}{\varepsilon} + \frac{64\Delta^2 R^2}{\varepsilon^2} \right\rceil \cdot \left\lceil \log_2 \left(1 + \frac{16\Delta^2}{\varepsilon L} \right) \right\rceil \tag{0.37}$$

вычислений субградиента f.

Замечание 0.0.22. Если не предполагать, что на (k+1)-й итерации ($k = 0, 1, \ldots, N-1$) модифицированного алгоритма 3 выполнено неравенство $L \leqslant L_{k+1} \leqslant 2L$ и предусмотреть полностью адаптивную настройку параметров L и Δ в методе, то оценка (3.19) может увеличиться не более, чем в

$$\left\lceil \max \left\{ \frac{2L}{L_0}, \frac{2\Delta}{\Delta_0} \right\} \right\rceil$$

раз. Отметим также, что логарифмический множитель в (3.19) можно опустить, если рассматривать неадаптивный вариант алгоритма 6 с фиксированным параметром $L_{k+1} = 2^p L$ при подходящем натуральном p.

В **разделе 3.4** предложен метод для вариационных неравенств с адаптивной настройкой величины, которая соответствует аддитивному шуму при задании оператора [175]. В частности, получен аналог теоремы 0.0.21 для вариационных неравенств. Обсуждаются результаты численных экспериментов, демонстрирующие эффективность разработанной процедуры для некоторых задач, в том числе для билинейных матричных игр с погрешностями.

В **главе 4** рассмотрены некоторые адаптивные методы для сильно выпуклых задач оптимизации, которые гарантируют линейную скорость сходимости в случае отсутствия погрешностей. Хорошо известно,

что в случае сильной выпуклости целевого функционала оценки скорости сходимости градиентного метода существенно улучшаются. Например, для сильно выпуклого целевого функционала с липшицевым градиентом известно, что градиентный метод сходится с линейной скоростью. Весьма популярен вопрос о том, насколько можно условие сильной выпуклости ослабить. В этом случае известен подход, основанный на использовании вместо сильной выпуклости условия градиентного доминирования Поляка-Лоясиевича ((PL)-условие) [45,114]

$$f(x) - f(x_*) \leqslant \frac{1}{2\mu} \|\nabla f(x)\|_2^2 \quad \forall x \in Q, \tag{0.38}$$

где x_* — точное решение задачи минимизации f. Известно, неравенство (0.38) в предположении липшицевости градиента f позволяет получить оценку скорости сходимости

$$\begin{aligned} f(x^N) - f(x_*) &\leqslant \left(1 - \frac{\mu}{L}\right)^N \left(f(x^0) - f(x_*)\right) \leqslant \\ &\leqslant \exp\left(-\frac{\mu}{L} N\right) \left(f(x^0) - f(x_*)\right). \end{aligned} \tag{0.39}$$

Отметим, что условие градиентного доминирования заведомо верно для сильно выпуклой целевой функции f относительно евклидовой нормы. Однако довольно хорошо известны примеры, когда нельзя быть уверенным даже в выпуклости $f(x)$, но (PL)-условие имеет место. Так, для записанной в векторном виде системы нелинейных уравнений $g(x) = 0$ (т.е. $g : \mathbb{R}^n \to \mathbb{R}^m$, $m \leqslant n$) рассмотрим задачу нахождения какого-нибудь решения этой системы. Если ввести матрицу Якоби отображения g : $\frac{\partial g(x)}{\partial x} = \left\| \frac{\partial g_i(x)}{\partial x_j} \right\|_{i,j=1}^{m,n}$ и предположить, что существует такое $\mu > 0$, что для всех $x \in \mathbb{R}^n$ имеет место равномерная невырожденность матрицы Якоби:

$$\lambda_{min}\left(\frac{\partial g(x)}{\partial x} \cdot \left[\frac{\partial g(x)}{\partial x}\right]^T\right) \geqslant \mu.$$

Тогда функция $f(x) = \|g(x)\|_2^2$ удовлетворяет условию (0.38) для произвольного x_* такого, что $f(x_*) = 0$, то есть $g(x_*) = 0$ [145].

В **разделе 4.1** введено следующее ослабление условия липшицевости градиента [52]

$$f(y) \leqslant f(x) + \langle \widetilde{\nabla} f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2 + \Delta \|y - x\|_2 \quad \forall x, y \in Q$$

для некоторых $L, \Delta > 0$. Например, это предположение естественно в случае, если $\widetilde{\nabla} f(x)$ — некоторое возмущенное с точностью Δ значение градиента $\nabla f(x)$, где f имеет L-липшицев градиент. По сути, левая часть неравенства (0.29) заменяется условием градиентного доминирования. Предложен следующий алгоритм 10 с адаптивным подбором шага с настройкой на величины L, Δ и показана оценка скорости сходимости, аналогичная (0.39).

Алгоритм 10 Адаптивный градиентный метод для функций, удовлетворяющих (PL)-условию.

Require: x^0 — начальная точка, параметры $\Delta_0,\ L_0$ $(2\mu \leqslant L_0 < 2L,\ \Delta_0 \leqslant 2\Delta)$.

1: $L_{k+1} := \max\{\mu, {}^{L_k}/_2\},\ \Delta_{k+1} := {}^{\Delta_k}/_2$.

2: $x^{k+1} = x^k - h_k \widetilde{\nabla} f(x^k)$,

$$h_k = \frac{1}{L_{k+1}} - \frac{\Delta_{k+1}}{L_{k+1}\widetilde{g}_{x^k}},\ \widetilde{g}_{x^k} = \left\|\widetilde{\nabla} f(x^k)\right\|_2.$$

3: **repeat**

4: **if** $f(x^{k+1}) \leqslant f(x^k) + \left\langle \widetilde{\nabla} f(x^k), x^{k+1} - x^k \right\rangle + \frac{L_{k+1}}{2}\left\|x^{k+1} - x^k\right\|_2^2 + \Delta_{k+1}\left\|x^{k+1} - x^k\right\|_2$ **then**

5: $k := k + 1$ и выполнение п. 1.

6: **else**

7: $L_{k+1} := 2 \cdot L_{k+1};\ \Delta_{k+1} := 2 \cdot \Delta_{k+1}$ и выполнение п. 2.

8: **end if**

9: **until** $k \geqslant N$

Для данного метода доказано, что либо невязка $\min\limits_k f(x^k) - f(x_*)$ убывает со скоростью геометрической прогрессии при увеличении k (см. (0.40)), либо эта невязка ограничена сопоставимой с Δ^2 величиной (это верно и если $h_k \leqslant 0$). Справедлива следующая

Теорема 0.0.23. *Пусть для некоторого натурального k $h_i > 0$ при $i \leqslant k$ и верно $\Delta_{k+1} \leqslant \Delta$. Тогда после k итераций алгоритма 10 для всякого $C > 1$ будет выполняться одно из двух неравенств*

$$f(x^{k+1}) - f(x_*) \leqslant \left(1 - \frac{\mu}{L}\left(\frac{C-1}{C+1}\right)^2\right)^{k+1}(f(x^0) - f(x_*)), \qquad (0.40)$$

где $\widehat{L} \leqslant 2L \max\left\{1, \frac{2\Delta}{\Delta_0}\right\}$ *или*

$$\min_{i=\overline{1,k+1}} f(x^i) - f^* < \frac{(C+1)^2\Delta^2}{2\mu}.$$

В **разделе 4.2** вводится аналог понятия (δ, L, μ)-оракула [91] для задач оптимизации, которая описывает класс задач, близких к гладким сильно выпуклым задачам оптимизации. Предложен метод с адаптивной настройкой параметра гладкости L, обоснована близкая к линейной скорость сходимости (с точностью до величины, определяемой погрешностями).

Определение 0.0.24. Будем говорить, что функция f допускает (δ, L, μ)-модель в точке x для некоторой константы $\mu > 0$, если для любого $y \in Q$ верно:

$$\mu V(y, x) \leqslant f(y) - (f_\delta(x) + \psi(y, x)) \leqslant LV(y, x) + \delta,$$

где $\psi(y, x)$ — выпуклая по y функция, $\psi(x, x) = 0$, $\delta > 0$.

Отметим, что обычное свойство μ-сильно выпуклости функции

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

здесь заменяется неравенством $\mu V(y, x) + f_\delta(x) + \psi(y, x) \leqslant f(x)$. При достаточно стандартном [138] условии на выбор прокс-функции обычная сильная выпуклость может гарантировать выполнение последнего неравенства.

Замечание 0.0.25. Пусть для всяких $x, y \in Q$ верно $d(y - x) \leqslant C_n\|y - x\|^2$, где n — размерность пространства и $C_n = O(\log n)$. Тогда $V(y, x) \leqslant C_n\|y - x\|^2$ и μC_n-сильная выпуклость $\frac{\mu}{2}||y - x||^2 \leqslant f(y) - f(x) - \psi(y, x)$ влечёт относительную μ-сильную выпуклость:

$$\mu V(y, x) + f_\delta(x) + \psi(y, x) \leqslant f(y).$$

В **разделе 4.3** по аналогии с определением 0.0.15 введена концепция (δ, Δ, L, μ)-модели целевой функции и предложен метод градиентного типа с адаптивной настройкой параметров неточности этой модели. По сути в этой части работы предложен метод с адаптивной настройкой параметров неточной модели на класс задач с условием, близким к

сильной выпуклости целевого функционала. Обоснована скорость сходимости этого метода, близкая к линейной (с точностью до величин, определяемых погрешностями).

В **разделе 4.4** предложен адаптивный метод [174] для задачи сильно выпуклой оптимизации с одним функциональным ограничением

$$f(x) \to \min, \quad x \in Q \subset \mathbb{R}^n,\ g(x) \leqslant 0, \tag{0.41}$$

где Q — выпуклый компакт в конечномерном нормированном пространстве $\mathbb{R}^n$, f и g — выпуклые функционалы, g удовлетворяет условию Липшица

$$|g(y) - g(x)| \leqslant M_g \|x - y\|_2 \ \ \forall x, y \in Q$$

при некоторой постоянной $M_g > 0$ ($\|\cdot\|_2$ — евклидова норма). При этом если функционалов ограничений несколько $\{g_p(x)\}_{p=1}^m$, то вполне можно рассмотреть задачу с одним ограничением $g(x) = \max\limits_{p=1,m} g_p(x)$, которое будет заведомо удовлетворять условию Липшица, если все g_p удовлетворяют условию Липшица. Вполне естественно рассматривать подход, основанный на замене (0.41) двойственной к ней задачей

$$\varphi(\lambda) = \min_{x \in Q}\{f(x) + \lambda g(x)\} \to \max_{\lambda \geqslant 0}. \tag{0.42}$$

В этом случае двойственная функция зависит от одной двойственной переменной $\lambda \geqslant 0$. Если выполнены условия Слейтера для задачи (0.41), то возможные значения λ ограничены отрезком. Это позволяет применять метод дихотомии для нахождения значения двойственной переменной λ, которое близко к соответствующему оптимальному значению λ_*, для которого

$$\lambda_* \cdot g(x(\lambda_*)) = 0.$$

Однако для эффективного решения (0.42) необходимо эффективно решать вспомогательную задачу многомерной минимизации по x функционала $f(x) + \lambda g(x)$ при фиксированном λ. Вообще говоря, такая задача может быть решена лишь с некоторой точностью методами оптимизации. Это приводит к погрешностям при нахождении $\varphi(\lambda)$ и ее производной $\varphi'(\lambda)$. Также если $f(x) + \lambda g(x)$ не сильно выпукла, то φ может быть негладкой в точке λ. Поэтому при рассмотрении указанного подхода вполне естественно потребовать сильную выпуклость целевого функционала. Исследовано влияние указанных погрешностей при решении

двойственной задачи на качество решения прямой задачи с учётом одномерности двойственной переменной, а также условий на гладкость и (сильную) выпуклость f или g. Предложен алгоритм с адаптивным критерием остановки вида

$$\lambda \cdot |g(x_\varepsilon(\lambda)| \leqslant \varepsilon \tag{0.43}$$

для задач вида (0.41), где $x_\varepsilon(\lambda)$ — ε-точное решение по функции вспомогательной задачи минимизации (0.42) при текущем двойственном множителе λ. Рассмотрен класс задач с сильно выпуклым целевым функционалом f при следующих типах предположений для f и g:

$$|f(x) - f(y)| \leqslant M_f \|x - y\|_2, \quad |g(x) - g(y)| \leqslant M_g \|x - y\|_2 \tag{0.44}$$

или

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L_f \|x - y\|_2, \quad \|\nabla g(x) - \nabla g(y)\| \leqslant L_g \|x - y\|_2 \tag{0.45}$$

для всех $x, y \in Q$ и для некоторых действительных положительных чисел M_f, M_g, L_f, L_g. Доказано, что при условиях (0.45) предлагаемый метод позволяет получить приемлемое качество решения задачи (0.41) после не более чем

$$O\left(\log_2^2 \frac{1}{\varepsilon}\right)$$

обращений к подпрограмме, выдающей градиент f или g. В предположениях (0.44) (т.е. для негладких задач) оптимальной оценки сложности предложенного подхода обосновать не удалось. Однако адаптивность предложенного критерия остановки (0.43) может позволить существенно повысить скорость его работы по сравнению с теоретическими оценками и для некоторых примеров негладких задач, что проиллюстрировано экспериментами.

В **разделе 4.5** предложен подход для задач выпуклого программирования с двумя функционалами ограничений, который аналогичен рассмотренному в разделе 4.4. В этом случае двойственная задача уже двумерна и вполне естественно применить к ней вариант метода минимизации выпуклой липшицевой функции двух переменных на квадрате с фиксированной стороной [44]. Идея метода — деление квадрата на меньшие части и постепенное их удаление так, чтобы в оставшейся достаточно малой части все значения целевой функции были достаточно близки к оптимальному. Показано, что метод может работать для

задач выпуклой гладкой оптимизации при наличии погрешностей решения вспомогательных одномерных задач, а также при вычислении направлений градиентов. Исследована возможная корреляция между указанными погрешностями. Также описана ситуация, когда возможно уменьшить временные затраты на решение вспомогательных одномерных задач. Данная методика приводит к подходу для решения двойственных задач к задаче многомерной сильно выпуклой минимизации с двумя выпуклыми функционалами ограничений. Этот подход основан на использовании для 2-мерной двойственной задачи предложенного метода Ю.Е. Нестерова с аналогичным (0.43) критерием остановки, который соответствует подходящему значению возмущенного градиента двойственной задачи.

Глава 5 посвящена новым вариантам схем зеркального спуска с переключениями для следующего класса задач оптимизации

$$f(x) \to \min \tag{0.46}$$

с M_g-липшицевым выпуклым функциональным ограничением $g(x) \leqslant 0$. Далее будем полагать, что x_* — одно из решений поставленной задачи (0.46). Предлагаются алгоритмические методы зеркального спуска с новыми адаптивными критериями остановки, выполнение которых гарантирует достижение приемлемого качества решения, например (с точностью умножения на некоторые постоянные): $f(\widehat{x}) - f(x_*) \leqslant \varepsilon$ и $g(\widehat{x}) \leqslant \varepsilon$. Использование адаптивных критериев позволяет ускорить работу методов, а также в некоторых случаях применять их и для задач, для которых не удаётся установить приемлемые оптимальные оценки сложности. В случае негладкости целевого функционала или функциональных ограничений естественно использовать субградиентные методы с переключениями, восходящие к хорошо известным работам [48, 62]. Зеркальные спуски для задач выпуклой оптимизации с ограничениями были предложены в [36]. В работе [68], предложены некоторые методы зеркального спуска для задач вида (0.46) с адаптивным выбором шагов и, самое важное, — с адаптивными критериями остановки. Глава 5 диссертации посвящена развитию некоторых идей указанной работы.

Через $(E, \|\cdot\|)$ обозначим конечномерное нормированное векторное пространство и E^* — сопряженное пространство к E со стандартной нормой $\|y\|_* = \max\limits_x \{\langle y, x\rangle, \|x\| \leq 1\}$, где $\langle y, x\rangle$ — значение линейного непрерывного функционала y в точке $x \in E$. Пусть $Q \subset E$ — некоторое

замкнутое выпуклое множество. Стандартно введём оператор проектирования

$$\mathrm{Mirr}_x(p) = \arg\min_{u \in Q} \left\{ \langle p, u \rangle + V(u, x) \right\} \text{ для всяких } x \in Q \text{ и } p \in E^*$$

и сделаем предположение о том, что оператор $\mathrm{Mirr}_x(p)$ легко вычислим.

В **разделе 5.1** выполнено теоретическое исследование метода

Алгоритм 17 Адаптивный зеркальный спуск, разные условия гладкости целевого функционала в случае одного ограничения.

if $g(x^N) \leqslant \varepsilon$ **then**
  $x^{N+1} = Mirr_{x^N}\left(\frac{\varepsilon \nabla f(x^N)}{\|\nabla f(x^N)\|_*}\right)$ // *"продуктивные шаги"* $(N \in I)$
else
  $x^{N+1} = Mirr_{x^N}\left(\frac{\varepsilon \nabla g(x^N)}{\|\nabla g(x^N)\|_*^2}\right)$ // *"непродуктивные шаги"*$(N \notin I)$
end if

Отметим, что алгоритм 17 применим к условным задачам выпуклой минимизации и позволяет получить оптимальные оценки скорости сходимости для достаточно широкого класса целевых функционалов в случае липшицева функционала ограничения. Например, в задачах с квадратичными целевыми функционалами мы сталкиваемся с ситуацией, когда такой функционал не удовлетворяет обычному свойству Липшица (или константа Липшица достаточно большая, если речь об ограниченном допустимом множестве), но градиент удовлетворяет условию Липшица. Можно рассматривать и более широкий класс уже негладких целевых функционалов

$$f(x) = \max_{1 \leqslant i \leqslant m} f_i(x), \text{ где} \tag{0.47}$$

$$f_i(x) = \frac{1}{2}\langle A_i x, x \rangle - \langle b_i, x \rangle + \alpha_i, \ i = 1, \ldots, m, \tag{0.48}$$

в случае, когда A_i $(i = 1, \ldots, m)$ — положительно определённые матрицы: $x^T A_i x \geqslant 0 \ \forall x \in Q$. Отметим, что функционалы вида (0.47)–(0.48) возникают в задачах проектирования механических конструкций Truss Topology Design со взвешенными балками. Эти функционалы, вообще говоря, не удовлетворяют условию Липшица.

Показано, как можно оценить скорость сходимости предлагаемого метода. Для всякого ненулевого конечного субградиента $\nabla f(x)$ целевого

функционала f рассматривается следующая величина

$$v_f(x, x_*) = \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|_*}, x - x_* \right\rangle, \quad x \in Q,$$

где x_* — искомое решение задачи (0.46). Случай $\nabla f(x) = 0$ здесь опускается, поскольку тогда x автоматически будет искомой точкой x_*. Для получения оценок полезно следующее известное вспомогательное утверждение [41].

Лемма 0.0.26. *Введем функцию* $\omega(\tau) = \max\limits_{x \in Q}\{f(x) - f(x_*) : \|x - x_*\| \leqslant \tau\}$, *где* $\tau > 0$. *Тогда для всякого* $y \in Q$ *верно неравенство*

$$f(y) - f(x_*) \leqslant \omega(v_f(y, x_*)).$$

Пусть x^0 — начальное приближение и постоянная Θ_0 такова, что $V(x_*, x^0) \leqslant \Theta_0^2$. В таком случае для алгоритма 17 справедлива следующая

Теорема 0.0.27. *Пусть* $\varepsilon > 0$ — *фиксированное число и выполнен критерий остановки*

$$\frac{2\Theta_0^2}{\varepsilon^2} \leqslant \sum_{k \notin I} \frac{1}{\|\nabla g(x^k)\|_*^2} + |I|$$

алгоритма 17. Тогда

$$\min_{k \in I} v_f(x^k, x_*) < \varepsilon.$$

Отметим, что алгоритм 17 работает не более

$$N = \left\lceil \frac{2 \max\{1, M_g^2\}\Theta_0^2}{\varepsilon^2} \right\rceil \tag{0.49}$$

итераций.

Хорошо известно, что оценка (0.49) оптимальна с точностью умножения на константу на классе задач выпуклого программирования с липшицевым целевым функционалом или ограничением. На базе леммы 0.0.26 и теоремы 0.0.27 можно, в частности, оценить скорость сходимости алгоритма 17 для следующего класса, вообще говоря, негладких целевых функционалов.

Следствие 0.0.28. *Пусть* $f(x) = \max\limits_{i=\overline{1,m}} f_i(x)$, *где* f_i *дифференцируема для всякого* $x \in Q$ *и* $\|\nabla f_i(x) - \nabla f_i(y)\|_* \leqslant L_i\|x-y\| \quad \forall x, y \in Q$. *Тогда после остановки алгоритма 17 верна оценка:*

$$\min_{k\in I} f(x^k) - f(x_*) \leqslant \varepsilon \cdot \|\nabla f(x_*)\|_* + \frac{L\varepsilon^2}{2} \ \textit{и} \ \max_{k\in I} g(x^k) \leqslant \varepsilon,$$

где $L = \max\limits_{i=\overline{1,m}} L_i$.

Доказано, что если задача (0.46) разрешима и её целевой функционал f удовлетворяет условию Гёльдера $|f(x) - f(y)| \leqslant M_\nu \|x-y\|^\nu \quad \forall x, y \in Q$ для некоторого $\nu \in [0;1)$, то для алгоритма 17 сохраняется оптимальная оценка скорости сходимости $O(\varepsilon^{-2})$ [53].

В **разделе 5.2** исследованы следующие две алгоритмические схемы с видоизменённым критерием проверки продуктивности шага, а также с новыми критериями остановки. Для проведённых экспериментов установлено, что такие методы могут работать быстрее по сравнению с предложенными в [68], если нормы субградиентов функционала ограничения могут быть большими.

Алгоритм 19 Адаптивный зеркальный спуск.

if $g(x^N) \leqslant \varepsilon\|\nabla g(x^N)\|_*$ **then**
  $x^{N+1} = Mirr_{x^N}\left(\frac{\varepsilon\nabla f(x^N)}{\|\nabla f(x^N)\|_*^2}\right)$ // *"продуктивные шаги"* $(N \in I)$
else
  $x^{N+1} = Mirr_{x^N}\left(\frac{\varepsilon\nabla g(x^N)}{\|\nabla g(x^N)\|_*}\right)$ // *"непродуктивные шаги"* $(N \notin I)$
end if

Ensure:

$$\widehat{x} = \sum_{k\in I} h_k x^k / \sum_{k\in I} h_k.$$

В виде таблицы 2 приведём сравнительный анализ предложенных алгоритмических схем [51, 176].

В **разделе 5.3** показано, что оценки скорости сходимости для алгоритмов 17 и 20 сохраняются в случае, если f — квазивыпуклый функционал. При этом в качестве аналога субградиента мы рассматриваем элементы $\nabla_{Cl} f(x)$ субдифференциала Кларка в предположении локальной липшицевости целевого функционала, а также $\nabla_{Cl} f(x) \neq 0$ при $x \neq x_*$.

Алгоритм 20 Адаптивный зеркальный спуск с фиксированным числом шагов.

1: **if** $g(x^N) \leqslant \varepsilon \|\nabla g(x^N)\|_*$ **then**
2: $\quad x^{N+1} = Mirr_{x^N}\left(\frac{\varepsilon \nabla f(x^N)}{\|\nabla f(x^N)\|_*}\right)$ // *"продуктивные шаги"* $(N \in I)$
3: **else**
4: $\quad x^{N+1} = Mirr_{x^N}\left(\frac{\varepsilon \nabla g(x^N)}{\|\nabla g(x^N)\|_*}\right)$ // *"непродуктивные шаги"* $(N \notin I)$
5: **end if**

Ensure: $\widehat{x} = \underset{x^k,\, k \in I}{\arg\min}\, f(x^k)$.

Таблица 0.1. Алгоритмы 19 – 20.

	Критерий остановки
Алгоритм 19	$2\frac{\Theta_0^2}{\varepsilon^2} \leqslant \sum\limits_{j \in I} \frac{1}{\|\nabla f(x^N)\|_*^2} + N - \lvert I \rvert$
Алгоритм 20	$2\frac{\Theta_0^2}{\varepsilon^2} \leqslant N$
	Оценки
Алгоритм 19	$f(\bar{x}^N) - f(x_*) \leqslant \varepsilon,\ \max\limits_{k \in I} g(x^k) \leqslant \varepsilon M_g$
Алгоритм 20	$\min\limits_{k \in I} v_f(x^k, x_*) \leqslant \varepsilon,\ \max\limits_{k \in I} g(x^k) \leqslant \varepsilon M_g$

В общем случае вместо $\nabla_{Cl} f(x)$ можно рассмотреть некоторый вектор нормали $Df(x) \in \widehat{D}f(x)$ ко множеству уровня f в точке x:

$$\widehat{D}f(x) = \{p \mid \langle p, x - y \rangle \geqslant 0 \ \ \forall y \in Q : \ f(y) < f(x)\}.$$

Вообще говоря, $\widehat{D}f(x)$ — непустой, замкнутый выпуклый конус. Следуя [41], мы предполагаем $\widehat{D}f(x) \neq \{0\}$ при $x \neq x_*$.

В **разделе 5.4** показано как можно предложить схемы зеркального спуска с переключениями, которые оптимальны для сильно выпуклых задач негладкой оптимизации. Рассмотрим задачу (0.46) для μ-сильно выпуклых функционалов f и g с одинаковым параметром $\mu > 0$. Несколько модифицируем предположения на прокс-функцию и допустим, что $d(x)$ ограничена на единичном шаре относительно выбранной нормы $\|\cdot\|$:

$$d(x) \leqslant \Theta_0^2, \quad \forall x \in Q : \|x\| \leqslant 1.$$

Также предположим, что для начальной точки $x^0 \in Q$ существует такое $R_0 > 0$, что $\|x_0 - x_*\|^2 \leqslant R_0^2$. Для построения методов решения силь-

но выпуклой задачи (0.46) использована идея рестартов (перезапусков) алгоритмов 17 – 20.

В качестве примера приведём оценку для следующего класса негладких сильно выпуклых целевых функционалов. Пусть

$$f(x) = \max_{i=\overline{1,m}} f_i(x), \tag{0.50}$$

где f_i дифференцируемы во всякой точке $x \in Q$ и имеют липшицев градиент, т.е. существуют $L_i > 0$ такие, что $\|\nabla f_i(x) - \nabla f_i(y)\|_* \leqslant L_i \|x - y\| \quad \forall x, y \in Q$. Рассмотрим функцию $\tau : \mathbb{R}^+ \to \mathbb{R}^+$:

$$\tau(\delta) = \max\left\{\delta \|\nabla f(x_*)\|_* + \frac{\delta^2 L}{2},\ \delta\right\}, \qquad \text{где } L = \max_{i=\overline{1,m}}\{L_i\}.$$

Ясно, что функция τ возрастает, $\tau(0) = 0$ и поэтому для всякого $\varepsilon > 0$ существует $\widehat{\varphi}(\varepsilon) > 0 : \ \tau(\widehat{\varphi}(\varepsilon)) = \varepsilon$. Рассмотрим [59] следующий метод и теоретический результат о его сходимости для задачи (0.46) при сделанных предположениях.

Алгоритм 23 Адаптивный алгоритм зеркального спуска для сильно выпуклых функционалов.

Require: точность $\varepsilon > 0$; начальная точка x^0;
Θ_0 s.t. $d(x) \leqslant \Theta_0^2 \quad \forall x \in Q : \|x\| \leqslant 1$; Q; $d(\cdot)$;
параметр сильной выпуклости μ; R_0, удовл. $\|x^0 - x_*\|^2 \leqslant R_0^2$.

1: Set $d_0(x) = d\left(\frac{x - x^0}{R_0}\right)$.
2: Set $p = 1$.
3: **repeat**
4: Set $R_p^2 = R_0^2 \cdot 2^{-p}$.
5: Set $\varepsilon_p = \frac{\mu R_p^2}{2}$.
6: Set x^p — выход алгоритма 17 с точностью $\widehat{\varphi}(\varepsilon_p)$, прокс-функцией $d_{p-1}(\cdot)$ и Θ_0^2.
7: $d_p(x) \leftarrow d\left(\frac{x - x^p}{R_p}\right)$.
8: Set $p = p + 1$.
9: **until** $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$.

Теорема 0.0.29. *Пусть f имеет вид (0.50), а также f и g — μ-сильно выпуклые функционалы на $Q \subset \mathbb{R}^n$ и $d(x) \leqslant \Theta_0^2$ для всех $x \in Q$ таких,*

что $\|x\| \leqslant 1$. *Предположим, что начальное приближение* $x^0 \in Q$ *и число* $R_0 > 0$ *заданы так, что* $\|x^0 - x_*\|^2 \leqslant R_0^2$. *Тогда при достаточно большом* $M_g > 0$ *для* $\widehat{p} = \left\lceil \log_2 \dfrac{\mu R_0^2}{2\varepsilon} \right\rceil$ *выход* $x_{\widehat{p}}$ *есть* ε*-решение задачи* (0.46), *а также верно неравенство*

$$\|x^{\widehat{p}} - x_*\|^2 \leqslant \frac{2\varepsilon}{\mu}.$$

При этом количество итераций алгоритма 17 при работе алгоритма 23 согласно пункту 6 листинга не превышает

$$\widehat{p} + \sum_{p=1}^{\widehat{p}} \frac{2\Theta_0^2 \max\{1, M_g^2\}}{\widehat{\varphi}^2(\varepsilon_p)}, \quad \text{где } \varepsilon_p = \frac{\mu R_0^2}{2^{p+1}}.$$

Замечание 0.0.30. Предыдущую оценку количества итераций работы алгоритма 17 при работе алгоритма 23 (рестарты) можно несколько конкретизировать при условии $\varepsilon < 1$. В этом случае при всяком $\delta < 1$ имеем $\tau(\delta) \leqslant C\delta$ для некоторой константы C. Поэтому можно считать, что $\widehat{\varphi}(\varepsilon) = \widehat{C} \cdot \varepsilon$ для соответствующей константы $\widehat{C} > 0$ и с точностью до умножения на константу имеем:

$$N \leqslant \widehat{p} + \frac{64\Theta_0^2 \max\{1, M_g^2\}}{\mu\varepsilon}.$$

В **разделе 5.5** описан адаптивный метод зеркального спуска для задач онлайн-оптимизации для выпуклых липшицевых целевых функционалов с липшицевым функциональным ограничением [180], обоснована оценка скорости сходимости и оптимальность метода с точки зрения известных нижних оценок на соответствующем классе задач.

В **разделе 5.6** обсуждаются приложения построенной теории к некоторым прикладным задачам. Например, показано, как предложенные методы зеркального спуска с переключениями могут быть применены к задаче распределения ресурсов, в частности к задаче оптимизации компьютерной сети. Уточним постановку задачи [50] [109]. Допустим, что имеется компьютерная сеть с n пользователями (узлами), которые обмениваются пакетами через фиксированный набор m соединений. Структура сети задана матрицей маршрутизации $C = (C_i^j) \in \mathbb{R}^{m \times n}$, столбцы которой $\mathbf{C}_i \neq 0$, $i = 1, \ldots, n$ есть булевы m-мерные векторы такие, что $C_i^j = 1$ в случае использования узлом i соединения j, в

противном случае $C_i^j = 0$. Ограничения на пропускную способность соединений задаются вектором $\mathbf{b} \in \mathbb{R}_+^m$ со строго положительными компонентами. Пользователи оценивают качество работы сети с помощью функций полезности $u_p(x_p)$, $p = 1, \ldots, n$, где $x_p \in \mathbb{R}_+^m$ — скорость передачи данных p–го пользователя. Согласно [117] в качестве критерия оптимальности системы можно принять сумму функций полезностей для всех пользователей и тогда задача максимизации суммарной полезности сети при заданных ограничениях на пропускную способность соединений формулируется следующим образом:

$$\max_{\left\{C\mathbf{x}=\sum\limits_{p=1}^{n}\mathbf{C}_p x_p\right\}\leqslant\mathbf{b}} \left\{U(\mathbf{x}) = \sum_{p=1}^{n} u_p(x_p)\right\},$$

где $\mathbf{x} = (x_1, \ldots, x_n)$. Решением данной задачи будет оптимальное распределение ресурсов x_*.

Отметим, что целевые функционалы тут могут быть самого разного уровня гладкости, в том числе даже формально не удовлетворяющие условию Липшица: $u_p(x) = \ln x_p$ или $u_p(x_p) = \sqrt{x_p}$. Предложенные субградиентные схемы с переключениями представляются более выгодными в случае большого количества соединений m, так как оценки скорости сходимости не зависят от размерности задачи. При этом показано, что алгоритмы 17 и 20 приводят к оптимальным оценкам сложности для гёльдеровых функций полезности. Для логарифмической функции полезности $u_p(x) = \ln x_p$ можно модифицировать допустимое множество задачи отступами от нуля по значениям переменных ($x_p \geqslant n\varepsilon$) и показать оценку сложности $O(\varepsilon^{-4})$ для алгоритмов 17, 19 и 20. Если же для некоторого $R > 0$ добавить условие ограниченности $\|x\|_2 \leqslant R$ и оценку сложности $O(\varepsilon^{-4})$ для алгоритмов 17 и 20 можно заменить на $O(n^4\varepsilon^{-2}\ln^4 n\varepsilon)$. Однако при этом оценка сложности зависит от величины, соответствующей количеству пользователей. Отметим при этом, что для проведённых экспериментов получилось, что алгоритм 19 за счёт адаптивного критерия остановки работает существенно быстрее по сравнению с алгоритмами 17 и 20.

Обсуждаются результаты численных экспериментов для других задач (в частности, аналоги задачи Ферма–Торричелли–Штейнера и задачи о наименьшем покрывающем шаре) с ограничениями по сравнению скоростей работы алгоритмов 17, 19 и 20 между собой, а также с

некоторыми аналогами [119,153]. По результатам проведённых экспериментов оказалось, что для алгоритма 19 за счёт предложенного адаптивного критерия остановки приемлемое качество решения достигается в несколько раз быстрее субградиентного метода [153].

В **разделе 5.7** рассмотрены приложения полученных результатов о зеркальных спусках к двум специальным классам задач. Во-первых, вместо условия Липшица для целевого функционала f и ограничения g рассмотрены неравенства

$$\langle \nabla f(x), x - y\rangle \leqslant M_f\sqrt{2V(y,x)} \text{ и } \langle \nabla g(x), x - y\rangle \leqslant M_g\sqrt{2V(y,x)}$$

для не обязательно 1-сильно выпуклой прокс-функции и дивергенции Брэгмана. Эти неравенства верны, в частности, в случае относительной липшицевости f (g) [125]. Рассмотрен частично адаптивный аналог алгоритма 19 при таких общих предположениях. Продуктивные шаги выбираются как $h_k^f = \frac{\varepsilon}{M_f^2}$, а непродуктивные — как $h_k^g = \frac{\varepsilon}{M_g}$. Продуктивность шага при этом проверяется с помощью неравенства $f(x^k) \leq \varepsilon M_g$. Введён критерий остановки метода $2V(x_*, x^0) \leqslant 2\Theta_0^2 \leqslant \frac{\varepsilon^2|I|}{M_f^2} + \varepsilon^2|J|$, после выполнения которого заведомо верны неравенства

$$f(\widehat{x}) - f(x_*) \leqslant \varepsilon \text{ и } g(\widehat{x}) \leqslant \varepsilon M_g,$$

где $\widehat{x} = \frac{1}{\sum\limits_{k\in I} h_k^f}\sum\limits_{k\in I} h_k^f x^k$. Отметим, что использование относительной липшицевости для функционала ограничения g позволяет расширить класс рассматриваемых задач по сравнению с упомянутыми ранее алгоритмами $17-20$, оптимальность оценок для которых обоснована в предыдущих разделах диссертации лишь в предположении липшицевости g.

Далее, рассмотрен класс задач выпуклой однородной оптимизации с относительной точностью при наличии функционалов ограничений. Такая постановка восходит к работам [150,151] (см. также главу 6 докторской диссертации Ю.Е. Нестерова [38]). Как показано в упомянутых выше работах, подход к оценке качества решения задачи в с точки зрения именно относительной точности вполне оправдан для разных прикладных задач (линейное программирование, проектирование механических конструкций и др.), если желаемая относительная точность искомого приближенного решения не очень велика. Известно, что достаточно широкий класс задач оптимизации с относительной точностью можно сводить к минимизации выпуклой (положительно) однородной

функции. Итак, рассматривается на множестве $Q \subset \mathbb{R}^n$ задача минимизации выпуклой однородной (мы используем термин «однородная» вслед за главой 6 из [38]) функции вида

$$f(x) \to \min_{x \in Q} \tag{0.51}$$

с выпуклыми функционалами ограничений $g_p(x) \leqslant 0, \quad p = \overline{1,m}$. Стандартно будем обозначать $g(x) := \max_{1 \leqslant p \leqslant m}\{g_p(x)\}$.

Во втором подпункте в предположении

$$\gamma_0\|x\| \leqslant f(x) \leqslant \gamma_1\|x\| \quad \forall\, x \in Q$$

для некоторых $\gamma_0 > 0$ и $\gamma_1 > 0$ показано, как субградиентный метод для двойственной задачи

$$f(x) + \lambda g(x) = f(x) + \sum_{p=1}^{m} \lambda_p g_p(x) \to \max_{\lambda \in \mathbb{R}^m},$$

где $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$. позволяет восстановить решение прямой задачи с относительной точностью по целевому функционалу

В разделе 5.7.3 описана возможность приложений разработанных адаптивных зеркальных спусков 19 и 20 к задачам выпуклого программирования с относительной точностью по целевому функционалу. Сначала приведена оценка скорости сходимости алгоритма 19 к выделенному классу задач [51] в случае, если 0 — внутренняя точка субдифференциала целевого функционала f в нуле.

Однако весьма важно, что также рассмотрен случай, когда 0 не обязательно есть внутренняя точка субдифференциала целевого функционала f в нуле. Рассмотрен следующий ослабленный вариант этого условия

$$B_{\gamma_0}^{K^*}(0) \subseteq \partial f(0) \subseteq B_{\gamma_1}^{K^*}(0),$$

где K^* — сопряженный конус к некоторому полунормированному конусу $K \subset \mathbb{R}^n$ с конус-полунормой $\|\cdot\|_K$ (отличие от обычной полунормы в том, что $\|\alpha x\|_K = \alpha\|x\|_K$ лишь для $\alpha \geqslant 0$), где под *сопряженным конусом* K^* понимается набор функционалов вида $\psi_\ell = \max\{0, \ell(x)\}$ для линейных функционалов $\ell : K \to \mathbb{R} : \ell(x) \leqslant C_\ell\|x\|_K$ при некотором $C_\ell > 0\ \forall\, x \in K$. Ясно, что K^* будет выпуклым конусом с операциями сложения $\psi_{\ell_1} \oplus \psi_{\ell_2} := \psi : \quad \psi(x) = \max\{0, \ell_1(x) + \ell_2(x)\}$ и умножения

на скаляр $\lambda \geqslant 0$ $\psi_{\lambda\ell}(x) = \lambda\psi_\ell(x) = \lambda \max\{0, \ell(x)\}$ $\forall\, x \in K$. На K^* можно ввести норму $\|\psi_\ell\|_{K^*} = \sup_{\|x\|_K \leqslant 1} \max\{0, \ell(x)\} = \sup_{\|x\|_K \leqslant 1} \ell(x)$ и шар $B_r^{K^*}(0) = \{\psi_\ell \in K^* \mid \|\psi_\ell\|_{K^*} \leqslant r\}$.

Из аналога теоремы об опорном функционале в нормированных конусах [171] получаем, что

$$\|x\|_K = \max_{\psi_\ell \in B_1^{K^*}(0)} \ell(x). \tag{0.52}$$

Отметим лишь, что при $\|x\|_K = 0$ достаточно в (0.52) выбрать $\ell \equiv 0$. Приведем некоторый пример пары (K, K^*).

Пример 0.0.31. Пусть $K = \{(x, y) \mid x, y \in \mathbb{R}\}$ и $\|(x, y)\|_K = \sqrt{x^2 + y^2} + y$. Можно проверить, что в таком случае

$$K^* = \{\psi_\ell \mid \ell((x, y)) = \lambda x + \mu y \,:\, \mu + \frac{\lambda^2}{\mu} < +\infty \text{ или } \lambda = \mu = 0\}, \text{ а}$$

$$\|\psi_\ell\|_{K^*} = \begin{cases} 0, & \text{если } \lambda = \mu = 0; \\ \dfrac{\mu}{2} + \dfrac{\lambda^2}{2\mu}, & \text{если } \mu + \dfrac{\lambda^2}{\mu} < +\infty \text{ при } \mu > 0. \end{cases}$$

Тогда $B_1^{K^*}(0)$ имеет вид круга на плоскости (λ, μ) радиуса 1 с центром в точке $\lambda = 0$, $\mu = 1$.

Не уменьшая общности рассуждений, будем полагать $K = \bigcup\limits_{r \geqslant 0} B_r^K(0)$, а также $x_* \in K$ для точного решения x_* рассматриваемой задачи минимизации f на Q. Для вывода оценок скорости сходимости методов с относительной точностью необходимо знать оценку R величины $\|x^0 - x_*\|_K$. Однако в выпуклых конусах, вообще говоря, не задана операция вычитания и поэтому в качестве аналога нормы разности можно использовать метрику $d^K(x^0, x_*)$, где

$$d^K(x, y) = \sup_{\|\psi_\ell\|_{K^*} \leqslant 1} |\psi_\ell(x) - \psi_\ell(y)|.$$

Некоторые условия, при которых нормированный конус допускает существование метрики такого типа, исследованы в [57, 173]. Получен аналог теоремы 6.1.1 [38] для указанного выше предположения. Будем также считать, что $x^0, x_* \in K$), причём $\|x^0\|_K = \min\{\|x\|_K,\ x \in Q\}$.

Теорема 0.0.32. 1) $\forall\, x \in K \quad \gamma_0\|x\|_K \leqslant f(x) \leqslant \gamma_1\|x\|_K$. *Более того,*

$$\frac{\gamma_0}{\gamma_1} f(x^0) \leqslant \gamma_0\|x^0\|_K \leqslant f^* \leqslant f(x^0) \leqslant \gamma_1\|x^0\|_K.$$

2) *Для всякого точного решения $x_* \in K$ справедливо неравенство:*

$$d^K(x^0, x_*) \leqslant \|x^0\|_K + \|x_*\|_K \leqslant \frac{2}{\gamma_0} f^* \leqslant \frac{2}{\gamma_0} f(x^0).$$

Ясно, что в случае $0 \in Q$ можно просто выбрать $x^0 = 0$, что позволит избежать необходимости дополнительно решать вспомогательную задачу минимизации нормы для выбора начальной точки x^0. Степень соблюдения ограничений задачи при этом предлагается проверять с помощью неравенств-проверки продуктивности итераций зеркального спуска. Заметим также, что для предыдущего результата существенен выбор начальной точки x^0. Иными словами, в случае $0 \notin Q$ важно быть уверенным в существовании точки, реализующей минимально возможное значение конус-полунормы $\|\cdot\|_K$ на множестве Q. Некоторые условия такого типа для множества Q получены в [56].

Далее, показана применимость к поставленной задаче выпуклой однородной минимизации приведенного выше алгоритма 19. Оказывается, что для этого достаточно выбрать прокс-структуру так, чтобы для некоторой константы $\hat{\omega} > 0$

$$V(x_*, x^0) \leqslant \hat{\omega} d^K(x_*, x^0). \tag{0.53}$$

Если верно (0.53) и $\hat{\omega} d^K(x_*, x^0) = \Theta_0^2 \geqslant 1$ (вместо метрики d^K теперь рассматривается $\hat{\omega} d^K$ с соответствующей константой $\gamma_0 > 0$), то для задачи (0.51) справедлива следующая

Теорема 0.0.33. *Пусть выпуклый однородный функционал f M_f-липшицев на Q. Тогда после $N \geqslant \frac{8 \max\{1, M_f^2\}}{\gamma_0^2 \delta^2}$ итераций предложенного аналога алгоритма 19 гарантированно будут выполняться неравенства:*

$$f(\widehat{x}) \leqslant f(x_*)(1+\delta) \text{ и } g(\widehat{x}) \leqslant \frac{M_g \Theta_0^2 \gamma_0 \delta}{2}.$$

Отметим, что использование метрики d^K для оценки качества решения здесь связано с отсутствием требования, чтобы 0 была внутренней точкой $\partial f(0)$. Для упрощения проверки применимости полученного результата в плане иных условий на f вполне можно использовать обычную евклидову прокс-структуру, подобрав подходящий параметр $\hat{\omega} > 0$ в (0.53).

Благодарности. Автор благодарит своих коллег и соавторов И. В. Орлова, А. В. Гасникова, П. Е. Двуреченского, А. И. Тюрина,

А. А. Титова, М. С. Алкуса, А. С. Иванову, А. Д. Агафонова, Д. А. Пасечнюка, а также И. В. Баран за постоянное сотрудничество и полезные обсуждения. Работа выполнена при поддержке грантов РФФИ 18-31-20005 мол-а-вед (введение, разделы 1.4, 2.1, 2.3, 3.1, 3.2, 4.2, 4.3, 5.7.2), РНФ 18-71-0048 (разделы 1.5.1, 2.2, 2.4, частично 2.5, 3.3, 3.4.1, 4.1, 4.5, 5.1, 5.2, 5.4 – 5.6, 5.7.2), РФФИ 18-29-03071 мк (разделы 1.1 – 1.3, 4.4, заключение), РНФ 18-71-10044 (частично — разделы 2.5 и 4.4), а также гранта Президента РФ для молодых российских ученых-кандидатов наук МК-15.2020.1 (разделы 5.7.1 и 5.7.3). Огромная признательность члену научных коллективов по гранту РНФ 18-71-0048 и гранту Президента МК-15.2020.1 А. Н. Степанову за значимую помощь в технической работе и разработке компьютерных программ для вычислительных экспериментов.

ГЛАВА 1

Обзор литературы. Некоторые вспомогательные сведения о классах оптимизационных задач, рассматриваемых в работе

1.1 О развитии теории методов выпуклой негладкой оптимизации

Задачи математического программирования встречаются в самых разных приложениях, и весьма актуальна проблема поиска эффективных алгоритмических процедур для их численного решения. Среди этих задач отдельно можно выделить различные классы задач выпуклой оптимизации. Выпуклые оптимизационные задачи составляют практически единственный класс задач оптимизации, которые допускают построение методов с глобальными скоростными характеристиками, приемлемыми для большинства практических приложений. Выпуклый анализ и выпуклая оптимизация основаны на солидном математическом фундаменте, разработанном в основном в первой половине 20-го столетия математиками Г. Минковским, К. Каратеодори, Э. Хелли, В. Фенхелем, А. Александровым и другими.

Многие выпуклые задачи являются негладкими, что привело к необходимости развития методов недифференцируемой оптимизации. Разработки в этой области начались в СССР в 60-е годы прошлого века. Такие методы предназначались прежде всего для решения задач линейного программирования большой размерности, возникавших при моделировании важнейших экономических, экологических и других реальных явлений. Для этого в работах отечественных математиков Б. Т. Поляка, В. Ф. Демьянова, Л. В. Васильева и В. Н. Малоземова, Н. З. Шора, Б. Н. Пшеничного были исследованы необходимые условия

экстремума некоторых классов недифференцируемых функций, а также предложены итеративные процедуры, обобщающие классические методы градиентного спуска. В США и странах Западной Европы методы недифференцируемой оптимизации начали активно изучаться в 70-е годы прошлого века. При этом отдельные работы в данном направлении, обусловленные преимущественно потребностями развития теории игр (которая используется при моделировании конкуренции в рыночной экономике), встречались и раньше. Основной областью применения методов недифференцируемой оптимизации было построение оптимальных значений целевых функций в задачах дискретной оптимизации. Наряду с численными алгоритмами исследовались свойства обобщённых градиентов (или субдифференциалов) и полученные на их основе условия оптимальности, в том числе и для невыпуклых функций. Позднее круг приложений негладкого анализа существенно расширился. Широко известны разработки в данной области Р. Рокафеллара, Ф. Кларка, Дж. Варги, Ф. Мишеля и Ж.-П. Пено, Ж.-П. Обена и И. Экланда , Д. Аусселя и др., А. Д. Дж Борвейна и К. Жу, а также многих других.

После выхода монографии Р. Т. Рокафеллара [49] центр развития выпуклого анализа окончательно сместился в сторону теории методов оптимизации [29]. Известно немало книг по выпуклому анализу, а также по его оптимизационным приложениям. Основные приоритетные результаты в этой области принадлежат отечественным ученым: А. С. Антипину, Ф. П. Васильеву, Е. Г. Гольштейну, В. Ф. Демьянову, Ю. Г. Евтушенко, Ю. М. Ермольеву, А. Д. Иоффе, Б. С. Мордуховичу, А. С. Немировскому, Ю. Е. Нестерову, Б. Т. Поляку, Р. А. Поляку, Б. Н. Пшеничному, А. М. Рубинову, В. М. Тихомирову, Л. Г. Хачияну, Н. З. Шору, Д. Б. Юдину и другим.

Проблема разработки эффективных алгоритмических процедур для оптимизационных задач с функционалами различного уровня гладкости довольно актуальна в приложениях. Отметим лишь некоторые примеры современных работ, где естественно возникают негладкие задачи. Так, негладкие задачи возникают при решении проблем проектирования механических конструкций (Truss Topology Design) [119, 153, 167]. Также, вообще говоря, негладкими будут задачи распределения ресурсов [117], сводящиеся к максимизации совокупной полезности производителей при совместном использовании имеющихся ресурсов, причём

в таком случае возможны и целевые функционалы, которые удовлетворяют более слабым условиям гладкости по сравнению с условием Липшица [133] (и не удовлетворяют условию Липшица). Как частный случай выделим задачу распределения ресурсов в компьютерных сетях, исследованную в недавних работах [26, 50]. В качестве примера интересных негладких оптимизационных задач можно отметить также и некоторые геометрические негладкие задачи оптимизации типа Ферма–Торричелли–Штейнера [130], задачи о наименьшем покрывающем шаре, Convex Feasibility Problem [70].

В качестве важного направления приложений негладкой выпуклой оптимизации можно выделить задачи многомерной оптимизации, которые возникают для разреженных задач [153]. Основная идея, позволяющая сводить задачу огромной размерности к негладкой оптимизационной задаче, заключается в эквивалентности системы неравенств

$$\langle a_k, x\rangle - b_k \leqslant 0, \quad k = 1, \ldots, m \quad (m \gg 1),$$

неравенству для одного негладкого функционала ограничения:

$$\max_{k=1,\ldots,m} \{\langle a_k, x\rangle - b_k\} \leqslant 0.$$

Опишем подробнее основные известные результаты о вычислительных гарантиях (оценках скорости сходимости) для задач (в общем случае негладкой) выпуклой оптимизации [36, 46, 75, 83, 140]. Мы начнем с задач небольшой размерности, когда возможна ситуация $N \geqslant n$, где n — размерность пространства, а N — количество вызовов оракула (например, число вычислений субградиента f в текущей точке). Рассмотрим задачу выпуклой оптимизации

$$f(x) \to \min_{x \in Q},$$

где Q — выпуклое компактное множество простой структуры. С использованием N обращений к оракулу для субградиента ставится задача нахождения такой точки x^N, для которой

$$f\left(x^N\right) - f(x_*) \leqslant \varepsilon, \tag{1.1}$$

где $f^* = f(x_*)$ — минимальное значение функции в (1.1), x_* — точное решение задачи (1.1). Нижняя и верхняя оценки требуемого количества обращений к оракулу (с точностью до порядка множителя, имеющего логарифмическую зависимость от некоторой характеристики допустимого множества Q) равны $N \sim n \ln(\Delta f/\varepsilon)$, где $\Delta f =$

$\sup\limits_{x,y\in Q}\{f(y)-f(x)\}$. Согласно указанной оценке сходится метод центров тяжести (метод внутренней точки) [123, 158]. При $n=1$ этот метод является простым бинарным поиском [80]. Однако при $n>1$ реализовать этот метод уже непросто. Это связано с тем, что сложность каждой итерации слишком высока, поскольку на каждой итерации требуется нахождение центра тяжести текущего допустимого множества [83]. Как хорошо известно, для метода эллипсоидов [36, 166] достижение приемлемого качества решения необходимо[1)] $N=\widetilde{O}\left(n^2\ln\left(\Delta f/\varepsilon\right)\right)$ вызовов оракула при сложности каждой итерации $O\left(n^2\right)$. В [83, 181] была предложена специальная версия метода отсекающих гиперплоскостей. Для такого метода требуется $N=\widetilde{O}\left(n\ln\left(\Delta f/\varepsilon\right)\right)$ вызовов оракула при сложности итерации $\widetilde{O}\left(n^{2.37}\right)$. В работе [122] предложен метод с $N=\widetilde{O}\left(n\ln\left(\Delta f/\varepsilon\right)\right)$ вызовами оракула и сложностью итерации $\widetilde{O}\left(n^2\right)$. Неочевидно, что этот метод практичный ввиду больших порядков логарифмов в $\widetilde{O}$.

Как видим, упомянутые выше оценки сложности существенно зависят от размерности задачи, что ставит под вопрос их применимость в случае большой размерности задачи n. Наиболее известный подход к пониманию эффективности алгоритмических процедур в многомерной оптимизации основан на теории сложности, восходящей к известной монографии А. С. Немировского и Д. Б. Юдина [36]. Разработанная А. С. Немировским и Д. Б. Юдиным теория верхних оценок возможной эффективности методов выпуклой минимизации для различных классов задач была подкреплена оптимальными методами, которые реализовывали эти оценки. Для класса методов, у которых на каждой итерации разрешается не более чем $\mathrm{O}(1)$ раз обращаться к оракулу (подпрограмме) для расчета градиента $\nabla f(x)$, оценка числа итераций N, необходимых для достижения точности решения задачи ε (по функции) (1.1) указана в таблице 1.1 в зависимости от класса рассматриваемых задач (x^N — выход работы метода). Как видим, указанные оценки сложности не зависят от размерности задачи.

Отметим, что в таблице 1.1 $R=\left\|x^0-x_*\right\|_2$ — расстояние от точки старта до точки-решения задачи x_*. Если решение не единственно, то в определении R под x_* можно понимать (евклидову) проекцию точки x^0 на множество решений. При этом оптимальные оценки достигаются для

[1)] Здесь и всюду далее для всех (больших) n: $\widetilde{O}(g(n))\leqslant C\cdot(\ln n)^r g(n)$ с некоторыми константами $C>0$ и $r\geqslant 0$. Как правило, $r=1$. Если $r=0$, то $\widetilde{O}(\cdot)=O(\cdot)$.

Таблица 1.1. Оптимальные оценки количества обращений к (суб)градиенту.

$N \leqslant n$	$\lvert f(y) - f(x)\rvert \leqslant$ $\leqslant M \lVert y - x\rVert$	$\lVert \nabla f(y) - \nabla f(x)\rVert_* \leqslant$ $\leqslant L \lVert y - x\rVert$
$f(x)$ выпукла	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$	$O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$
$f(x)$ μ-сильно выпукла в $\lVert\cdot\rVert$-норме	$O\left(\frac{M^2}{\mu\varepsilon}\right)$	$O\left(\sqrt{\frac{L}{\mu}}\left\lceil \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil\right)$ $(\forall N)$

хорошо известного быстрого градиентного метода Ю. Е. Нестерова [40].

После этого на некоторый период градиентные методы были практически забыты. Начиная с выдающейся работы Н. Кармаркара [115] и примерно до 2000 г. развитие теории и методов оптимизации было в основном связано с прогрессом в теории полиномиальных методов внутренней точки. К примеру, была разработана общая теория самосогласованных функций [147], которая позволяла строить полиномиальные методы внутренней точки для всех выпуклых задач с явной структурой. Однако, как было отмечено выше, сложность методов внутренней точки растет сильно зависит от размерности решаемой задачи. В связи с указанным обстоятельством в 2000-е годы возник интерес к оптимизационным методам градиентного типа, итерации которых требуют меньших затрат памяти, что интересно ввиду приложений к задачам оптимизации с большими данными. Тогда вновь возник интерес к теории сложности [36] и подкрепляющим её методам градиентного типа.

Таблица 1.1 описывает (более подробно см. [75,83,140]) оптимальные оценки числа вызовов оракула для задачи выпуклой оптимизации (1.1) вне зависимости от размерности задачи n. Здесь уже Q — необязательно компактное множество. В таблице 1.1 R — это "расстояние"(с точностью до множителя $\ln n$) между начальной точкой и ближайшим решением $R = \widetilde{O}\left(\left\|x^0 - x_*\right\|\right)$.

Опишем оптимальный метод (для которого достигаются указанные границы сложности) на классе выпуклых липшицевых целевых функционалов в самом простом случае: $Q = \mathbb{R}^n$ и $\|\cdot\| = \|\cdot\|_2$ (см. [46,148]). В этом случае оптимальным будет обычный субградиентный метод. Пусть $B_2^n(x_*, R) = \{x \in \mathbb{R}^n : \ \|x - x_*\|_2 \leqslant R\}$. Будем всюду далее обозначать субградиент f (некоторый элемент субдифференциала $\partial f(x)$) в точке x как $\nabla f(x)$. Если f дифференцируема в точке x, то $\nabla f(x)$ — ее градиент.

Итерационный процесс (суб)градиентного метода имеет вид

$$x^{k+1} = x^k - h\nabla f\left(x^k\right). \tag{1.2}$$

Предположим, что для $x \in B_2^n\left(x_*, \sqrt{2}R\right)$

$$\left\|\nabla f\left(x\right)\right\|_2 \leqslant M,$$

где $R = \left\|x^0 - x_*\right\|_2$. Поэтому

$$\left\|x - x^{k+1}\right\|_2^2 = \left\|x - x^k + h\nabla f\left(x^k\right)\right\|_2^2 =$$

$$= \left\|x - x^k\right\|_2^2 + 2h\left\langle \nabla f\left(x^k\right), x - x^k\right\rangle + h^2\left\|\nabla f\left(x^k\right)\right\|_2^2 \leqslant$$

$$\leqslant \left\|x - x^k\right\|_2^2 + 2h\left\langle \nabla f\left(x^k\right), x - x^k\right\rangle + h^2M^2. \tag{1.3}$$

Выберем теперь $x = x_*$ (если x_* не единственно, то выбирается ближайшее решение x_* к точке старта x^0). Имеем:

$$f\left(\frac{1}{N}\sum_{k=0}^{N-1} x^k\right) - f(x_*) \leqslant$$

$$\leqslant \frac{1}{N}\sum_{k=0}^{N-1} f\left(x^k\right) - f\left(x_*\right) \leqslant \frac{1}{N}\sum_{k=0}^{N-1}\left\langle \nabla f\left(x^k\right), x^k - x_*\right\rangle \leqslant$$

$$\leqslant \frac{1}{2hN}\sum_{k=0}^{N-1}\left(\left\|x_* - x^k\right\|_2^2 - \left\|x_* - x^{k+1}\right\|_2^2\right) + \frac{hM^2}{2} =$$

$$= \frac{1}{2hN}\left(\left\|x_* - x^0\right\|_2^2 - \left\|x_* - x^N\right\|_2^2\right) + \frac{hM^2}{2}.$$

Если выбрать шаг градиентного метода и выход следующим образом:

$$h = \frac{R}{M\sqrt{N}}, \quad \bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k, \tag{1.4}$$

то будет выполняться неравенство

$$f\left(\bar{x}^N\right) - f(x_*) \leqslant \frac{MR}{\sqrt{N}}. \tag{1.5}$$

Заметим, что точная нижняя оценка для фиксированного шага в методе первого порядка для класса задач выпуклой оптимизации с условием (1.1) [93]:

$$f\left(x^N\right) - f(x_*) \geqslant \frac{MR}{\sqrt{N+1}}.$$

Неравенство (1.5) означает, что при выборе

$$N = \frac{M^2R^2}{\varepsilon^2}, \quad h = \frac{\varepsilon}{M^2}$$

будет достигаться оценка $f\left(\bar{x}^N\right) - f(x_*) \leqslant \varepsilon$, которая оптимальна с точностью до умножения на константу (см. таблицу 1.1). Далее ввиду (1.2) при выборе в (суб)градиентном методе переменной величины шага

$$x^{k+1} = x^k - h_k \nabla f\left(x^k\right), \quad h_k = \frac{\varepsilon}{\left\|\nabla f\left(x^k\right)\right\|_2^2}$$

оценка (1.5) сохранится для выходной точки

$$\bar{x}^N = \frac{1}{\sum\limits_{k=0}^{N-1} h_k} \sum_{k=0}^{N-1} h_k x^k.$$

Если же в 1.1) выбрать шаги h_k по аналогии с (1.4):

$$h_k = \frac{R}{\left\|\nabla f\left(x^k\right)\right\|_2 \sqrt{N}},$$

то будет верна следующая оценка, похожая на (1.5):

$$\min_{k=0,\ldots,N-1} f\left(x^k\right) - f(x_*) \leqslant \frac{MR}{\sqrt{N}}$$

как для выпуклых, так и для квазивыпуклых (или унимодальных) функционалов f [42, 47]:

$$f(\alpha x + (1-\alpha)y) \leqslant \max\left\{f(x), f(y)\right\} \forall x, y \in Q, \alpha \in [0,1].$$

Теперь покажем, как можно построить метод, оптимальный на классе задач с выпуклыми липшицевыми функционалами в случае произвольной нормы $\|\cdot\|$. Для оценки качества решения можно применять дивергенцию Брегмана V, определение которой уточнялось в предыдущем разделе.

Обозначим через $R^2 = V(x_*, x^0)$, где x_* — решение (1.1) (если x_* не единственно, то полагаем, что x_* — ближайшее к x^0 с точки зрения соответствия дивергенции Брегмана). Естественное обобщение итерационного процесса (1.2) — следующий алгоритм зеркального спуска [37, 75]:

$$x^{k+1}=\mathrm{Mirr}_{x^k}\left(h\nabla f\left(x^k\right)\right),\ \text{где}$$

$$\mathrm{Mirr}_{x^k}(x) = \arg\min_{x\in Q}\left\{\left\langle h\nabla f(x^k), x - x^k\right\rangle + V(x, x^k)\right\}. \tag{1.6}$$

Отметим, что здесь уже Q — это в общем случае не все пространство. В этом случае речь уже идет о методе проекции субградиента. Для этого итерационного процесса вместо (1.3) выполняется неравенство

$$2V(x, x^{k+1}) \leqslant 2V\left(x, x^k\right) + 2h\left\langle \nabla f\left(x^k\right), x - x^k\right\rangle + h^2M^2,$$

где $\left\|\nabla f\left(x\right)\right\|_* \leqslant M$ для всякого $x : V(x_*, x) \leqslant 2V(x_*, x^0) = 2R^2$.

При таких предположениях справедливы следующие аналоги (1.4) и (1.5)

$$f\left(\bar{x}^N\right) - f(x_*) \leqslant \frac{\sqrt{2}MR}{\sqrt{N}},$$

где

$$\bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k, \quad h = \frac{\varepsilon}{M^2}.$$

В [75] показано, как возможно выбрать прокс-функцию $d(x)$ для различных допустимых выпуклых множеств Q. Заметим, что, как правило, можно гарантировать, что [75]

$$R \leqslant C\sqrt{\ln n}\cdot\left\|x_* - x^0\right\|.$$

Рассмотрим некоторые примеры.

Пример 1.1.1. Если $Q = \mathbb{R}^n$ и норма евклидова $\|\cdot\| = \|\cdot\|_2$, то естественно выбрать $d(x) = \frac{1}{2}\|x\|_2^2$ и $V(y, x) = \frac{1}{2}\|y - x\|_2^2$. В таком случае метод

$$x^{k+1} = \mathrm{Mirr}_{x^k}\left(h\nabla f\left(x^k\right)\right) =$$

$$= \arg\min_{x\in\mathbb{R}^n}\left\{h\left\langle\nabla f\left(x^k\right), x - x^k\right\rangle + \frac{1}{2}\|x - x^k\|_2^2\right\} = x^k - h\nabla f\left(x^k\right)$$

соответствует стандартному итерационному процессу (1.2).

Пример 1.1.2. Рассмотрим теперь задачу оптимизации на единичном симплексе. Иными словами, предположим, что допустимое задачи множество имеет вид

$$Q = S_n(1) = \left\{ x \in \mathrm{R}^n_+ : \sum_{i=1}^n x_i = 1 \right\}, \quad \|\nabla f(x)\|_\infty \leqslant M_\infty \quad \forall x \in Q.$$

Выберем ℓ_1-норму $\|\cdot\| = \|\cdot\|_1$, а также прокс-функцию, шаг и начальное приближений следующим образом

$$d(x) = \ln n + \sum_{i=1}^n x_i \ln x_i, \; h = M_\infty^{-1}\sqrt{2\ln n/N}, \quad x_i^0 = 1/n, \quad i = 1, ..., n. \tag{1.7}$$

В таком случае для всяких $k = 0, ..., N-1$, а также $i = 1, ..., n$ итерации зеркального спуска будут иметь следующий вид

$$x_i^{k+1} = \frac{\exp\left(-h\sum\limits_{r=1}^k \nabla_i f(x^r)\right)}{\sum\limits_{l=1}^n \exp\left(-h\sum\limits_{r=1}^k \nabla_l f(x^r)\right)} = \frac{x_i^k \exp\left(-h\nabla_i f\left(x^k\right)\right)}{\sum\limits_{l=1}^n x_l^k \exp\left(-h\nabla_l f\left(x^k\right)\right)}.$$

В таком случае оценку скорости сходимости можно выписать так:

$$f\left(\bar{x}^N\right) - f(x_*) \leqslant M_\infty\sqrt{\frac{2\ln n}{N}}, \quad \bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k.$$

Если же для задачи на симплексе вместо (1.7) использовать евклидову $\|\cdot\|_2$-норму и соответствующую ей прокс-функцию $d(x) = \frac{1}{2}\left\|x - x^0\right\|_2^2$, то стоимость итерации проектирования повысится и оценка скорости сходимости будет иметь вид

$$f\left(\bar{x}^N\right) - f(x_*) \leqslant \frac{M_2}{\sqrt{N}}, \quad \|\nabla f(x)\|_2 \leqslant M_2 \quad \forall x \in Q.$$

В этом случае $M_2 = \mathrm{O}\left(\sqrt{n}M_\infty\right)$, и поэтому для задач на симплексе более разумно использовать $\|\cdot\|_1$-норму.

Предположим теперь, что целевой функционал f в (1.1) дополнительно μ-сильно выпуклый относительно евклидовой $\|\cdot\|_2$-нормы:

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2 \quad \forall x, y \in Q.$$

Тогда рассмотрим градиентный метод

$$x^{k+1} = \mathrm{Mirr}_{x^k}\left(h_k \nabla f\left(x^k\right)\right) =$$

$$= \arg\min_{x \in Q}\left\{h_k \left\langle \nabla f\left(x^k\right), x - x^k\right\rangle + \frac{1}{2}\left\|x - x^k\right\|_2^2\right\},$$

где

$$h_k = \frac{2}{\mu \cdot (k+1)}, \quad d(x) = \frac{1}{2}\left\|x - x^0\right\|_2^2, \quad \left\|\nabla f(x)\right\|_2 \leqslant M \quad \forall x \in Q.$$

В таком случае получим следующую оценку качества решения [118]

$$f\left(\sum_{k=1}^{N} \frac{2k}{N(N+1)} x^k\right) - f(x_*) \leqslant \frac{2M^2}{\mu \cdot (N+1)},$$

оптимальную на классе задач оптимизации липшицевых сильно выпуклых целевых функционалов (см. таблицу 1.1).

Итак, методы градиентного типа достаточно просты и на классе липшицевых (возможно, и негладких) целевых функционалов приводят к оптимальным оценкам скорости сходимости. Однако полученные оценки эффективности для задач негладкой оптимизации представляются довольно пессимистичными. Сложность задачи минимизации выпуклого липшицева (в частности, и негладкого) функционала достигает $O\left(\varepsilon^{-2}\right)$ итераций градиентного типа, а для задачи минимизации выпуклого функционала с липшицевым градиентом — $O\left(\sqrt{\varepsilon^{-1}}\right)$. В этой связи естественно возникает вопрос о том, насколько возможно приблизить скорость сходимости численных методов для задач негладкой оптимизации к гладкому случаю. В этом плане хорошо известен подход, основанный на технике сглаживания [149]. Однако этот подход существенно связан со специально подобранной структурой рассматриваемых задач, которая позволяет подобрать удачную аппроксимацию негладкой целевой функции некоторым гладким сильно выпуклым аналогом. С другой стороны, даже в случае применимости техника сглаживания может не приводить к удовлетворительным результатам [119].

Поэтому вполне естественны дальнейшие усилия по нахождению новых подходов к ускорению работы алгоритмических методов для негладких оптимизационных задач. Среди них можно выделить так называемые *универсальные методы*, исследованию которых было положено начало в работе Ю. Е. Нестерова [152]. Указанный подход основан на построении для задач выпуклой оптимизации с гёльдеровым

(суб)градиентом целевого функционала (параметр $\nu \in [0;1]$ фиксирован)

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L_\nu \|x - y\|^\nu \quad \forall x, y \in Q,$$

аналога стандартной квадратичной интерполяции с искусственной введённой неточностью $\delta > 0$

$$f(x) + \langle \nabla f(x), y - x \rangle \leqslant f(y) \leqslant$$

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

где

$$L = L_\nu \left[\frac{L_\nu}{2\delta} \cdot \frac{1-\nu}{1+\mu} \right]^{\frac{1-\nu}{1+\nu}}.$$

Задачи с гёльдеровым градиентом могут возникать в практических ситуациях (например, в задаче оптимизации газотранспортных систем). При этом $\nu = 1$ означает, что градиент удовлетворяет свойству Липшица, а $\nu = 0$ — сама функция удовлетворяет свойству Липшица. Универсальность метода при этом понимается как возможность адаптивной настройки при работе метода на оптимальный в некотором смысле уровень гладкости задачи и величину, соответствующую константе Гёльдера L_ν градиента (возможно, субградиента при $\nu = 0$) целевого функционала. Оказывается, что возможность такой настройки может позволить для некоторых задач улучшить скорость сходимости по сравнению с нижними теоретическими оценками на классе задач.

Как известно, погрешности при решении задач оптимизации возникают в силу разных причин [46]. Они могут быть естественно связаны с неточностью доступных данных, заменой бесконечномерной задачи конечномерным аналогом и т. д. Возможны и искусственные неточности, возникающие в ходе математического исследования рассматриваемых задач. Помимо указанной выше идеологии универсальных методов в этом плане можно отметить и неточности, связанные с техникой регуляризации задачи, а также со сглаживанием [149].

Поэтому естественно возникает проблема описания влияния погрешностей задания целевого функционала и градиента на оценки скорости сходимости методов. Для градиентных методов выпуклой оптимизации известен подход, основанный на недавно предложенной концепции неточного оракула [90, 91]. Известно, что для неускоренных градиентных методов в оценках не происходит накопления величин, связанных

с погрешностями. Однако для оптимальных при отсутствии погрешностей на классе гладких задач ускоренных методов в итоговой оценке скорости сходимости величины погрешностей могут накапливаться. Известны подходы к этой проблеме для специального понятия детерминированного шума (погрешности) [89] при задании градиента, а также для случайного аддитивного шума (погрешности) [16, 88, 102] при задании градиента.

1.2 Относительная гладкость и относительная сильная выпуклость в задачах оптимизации

В основу получения оценок скорости сходимости для градиентного метода может быть положена идея аппроксимации функции в исходной точке (текущем положении метода) мажорирующим ее параболоидом вращения. Так, для задачи минимизации выпуклого функционала $f : Q \to \mathbb{R}$ с липшицевым градиентом ($L \geqslant 0$)

$$\|\nabla f(x) - \nabla f(y)\|_* \leqslant L \|x - y\| \quad \forall x, y \in Q$$

выполняются неравенства

$$f(x) + \langle \nabla f(x), y - x \rangle \leqslant f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 . \quad (1.8)$$

Неравенства (1.8) позволяют получить для обычного градиентного спуска оценку скорости сходимости

$$f(\widehat{x}) - f^* \leqslant \frac{C_1}{N},$$

где $\widehat{x}$ — выход работы метода после N итераций, f^* — точное значение искомого минимума функции f, C_1 — некоторая постоянная.

Недавно предложенное условие относительной гладкости [67], связанное с заменой стандартного неравенства (1.8) на ослабленный вариант

$$f(x) + \langle \nabla f(x), y - x \rangle \leqslant f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + LV(y, x).$$

предполагает лишь выпуклость (но не сильную выпуклость) порождающей функции d. Концепция относительной гладкости позволяет применить вариант градиентного метода для некоторых задач, которые ранее решались лишь с помощью методов внутренней точки. В частности, речь идет об известной задаче построения оптимального эллипсоида, покрывающего заданный набор точек, которая имеет приложения в статистике и анализе данных [124].

Опишем понятие относительной гладкости более подробно. По-прежнему, рассматривается задача

$$f^* := \min_{x \in Q} f(x), \tag{1.9}$$

где $Q \subseteq E$ — замкнутое выпуклое множество в конечномерном векторном пространстве E со скалярным произведением $\langle \cdot, \cdot \rangle$, а $f : Q \to \mathbb{R}$ — дифференцируемая выпуклая функция.

В настоящее время существует множество методов первого порядка для задач оптимизации (1.9) (см., например, [83, 87, 89]). Практически все такие методы применимы к задачам (1.9) в случае, когда градиент f удовлетворяет условию Липшица на Q, а именно существует константа $0 \leqslant L < \infty$, для которой:

$$||\nabla f(x) - \nabla f(y)||_* \leqslant L\|x - y\| \quad \forall\, x, y \in Q, \tag{1.10}$$

где $\|\cdot\|$ — исходная норма в $\mathbb{E}$, а $\|\cdot\|_*$ — обычная сопряженная (двойственная) норма. Если, к примеру, рассмотреть стандартную схему градиентного спуска в случае евклидовой нормы в (1.10):

$$x^{i+1} \leftarrow \arg\min_{x \in Q} \left\{ f(x^i) + \langle \nabla f(x^i), x - x^i \rangle + \frac{L}{2}\|x - x^i\|_2^2 \right\},$$

то хорошо известно, что после завершения k итераций ($i = 0, 1, 2, \ldots$) для всякого $x \in Q$ выполняется неравенство:

$$f(x^k) - f(x) \leqslant \frac{L\|x - x^0\|_2^2}{2N}. \tag{1.11}$$

Эта оценка означает сублинейную скорость сходимости рассматриваемого метода, сопоставимую с $O(1/N)$ [83, 87]. Если же, кроме этого, f еще и μ-сильно выпукла при некотором $\mu > 0$:

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2 \quad \forall\, x, y \in Q,$$

то можно доказать и линейную сходимость для градиентного метода [83, 87]. Точнее говоря, для любого $x \in Q$ будет верно неравенство

$$f(x^N) - f(x) \leqslant \frac{L}{2}\left(1 - \frac{2\mu}{L+\mu}\right)^N \|x - x^0\|_2^2.$$

Более общие версии методов первого порядка не ограничиваются евклидовой ($||\cdot||_2$) нормой и используют дифференцируемые прокс-функции d. Стандартная схема методов первого порядка для дивергенции Брегмана [89] имеет следующий вид:

$$x^{i+1} \leftarrow \arg\min_{x \in Q} \left\{ f(x^i) + \langle \nabla f(x^i), x - x^i \rangle + LV(x, x^i) \right\}. \tag{1.12}$$

Обратим внимание, что шаг (1.12) приводит к необходимости решать вспомогательные подзадачи вида:

$$x_{\text{new}} \leftarrow \arg\min_{x \in Q} \{\langle c, x \rangle + d(x)\} \tag{1.13}$$

для соответствующих c. Действительно, (1.13) является частным случаем подзадачи (1.13) при $c = \frac{1}{L}\nabla f(x^i) - \nabla d(x^i)$ на i-ой итерации. Особенно важно отметить, что схема градиентного метода несколько бессмысленна, если нет возможности эффективно решить подзадачу (1.12). При реализации метода первого порядка для задачи (1.9), как правило, пытаются указать норму $\|\cdot\|$ и соответствующую сильно выпуклую прокс-функцию d с учетом вида допустимой области Q. При этом делается предположение об эффективной разрешимости подзадачи (1.13). Что же касается вычислительных гарантий, то для схемы градиентного метода можно показать, что после завершения N итераций для любого $x \in Q$ заведомо верна оценка:

$$f(x^N) - f(x) \leqslant \frac{LV(x, x^0)}{N},$$

что является прямым обобщением (1.11) [89, 164].

Как правило, при использовании стандартных методов первого порядка для (1.9) требуется, чтобы функционал f имел L-липшицев градиент на Q. Это предположение обеспечивает выполнение необходимых оценок сложности. Однако на практике существует множество выпуклых дифференцируемых функций, которые не удовлетворяют условию Липшица для градиента.

Например, рассмотрим целевой функционал $f(x) := -\ln\det(HXH^T)$ с диагональной матрицей $X := \mathrm{Diag(x)}$ на множестве $Q = \{x \in \mathbb{R}^n : \langle e, x\rangle = 1,\ x \geqslant 0\}$ (e — орты), или $f(x) = |x|^3$, или $f(x) = x^4$ на допустимом множестве $Q = \mathbb{R}$, или $f(x) = -\ln(x) + x^2$ на $Q = \mathbb{R}_{++}$. Конечно, если предположить, итерации алгоритма гарантируют монотонное убывание значений целевого функционала (что верно для большинства методов первого порядка на классах гладких функций), то этого вполне достаточно для того, чтобы f было гладким на некотором множестве уровня функции f. Однако даже в этом случае постоянная L может быть очень большой. Например, при $f(x) = -\ln(x) + x^2$ на $Q = \mathbb{R}_{++}$ ($x > 0$) можно рассмотреть множество уровня $\{x : f(x) \leqslant 10\}$. Тогда $L \approx \exp^{20}$, что представляется нецелесообразным для практического использования.

Именно с целью преодоления указанных выше недостатков некоторое время назад введены условия *относительной гладкости* [67] и *относительной сильной выпуклости* целевого функционала относительно заданной прокс-функции d [124]. Такой подход не требует спецификации какой-либо конкретной нормы, и d не обязательно строго или сильно выпукла. Концепции относительной гладкости и относительной сильной выпуклости позволяют применять методы градиентного типа для решения более общего класса задач выпуклой оптимизации (без условия липшицевости градиента), а также приводить к результатам линейной сходимости для соответствующих классов функционалов.

Теперь напомним концепции относительной гладкости [67] и относительной сильной выпуклости целевого функционала [124]. Пусть d — произвольная дифференцируемая выпуклая функция (она необязательно должна быть сильно или даже строго выпуклой), определенная на Q. Всюду далее $\operatorname{int} Q$ означает внутренность множества Q. Если множество Q не имеет внутренних точек, то вместо внутренности можно использовать относительную внутренность Q.

Определение 1.2.1. Функция f называется L-гладкой относительно d на Q, если для любых $x, y \in \operatorname{int} Q$ существует постоянная $L > 0$:

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + LV(y, x) \quad \forall x, y \in Q. \tag{1.14}$$

Определение 1.2.2. Говорят, что f μ-сильно выпукла относительно d на Q, если существует постоянная $\mu \geqslant 0$, для которой

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle + \mu V(y, x) \quad x, y \in Q. \tag{1.15}$$

Отметим, что относительная гладкость и относительная сильная выпуклость f определяются непосредственно; в определениях нет нормы, так что гладкость/сильная выпуклость не зависят ни от какой нормы. Более того, предполагается, что d не обладает какими-либо особыми свойствами, такими как строгая или обычная сильная выпуклость, а ключевые структурные свойства включают поведение f относительно прокс-функции d. Приведенное выше определение относительной гладкости эквивалентно условию (LC) в [73], но в [73] ещё требуется, чтобы функция d была гладкой и строго выпуклой.

Приведем эквивалентные варианты определений относительной гладкости и относительной сильной выпуклости целевой функции [124]. В случае, когда f и d дважды дифференцируемы, утверждения (a-iii) и (b-iii) предложения демонстрируют, что приведенные выше условия (1.14) и (1.15) верны в случае

$$\mu\nabla^2 d(x) \preceq \nabla^2 f(x) \preceq L\nabla^2 d(x).$$

Предложение 1.2.3. *Следующие условия эквивалентны:*

(a-i) f *— L-гладкая относительно d,*

(a-ii) $Ld - f$ *— выпуклая функция на Q,*

(a-iii) Для дважды дифференцируемых f и d верно $\nabla^2 f(x) \preceq L\nabla^2 d(x)$ *для любого* $x \in int\, Q$,

(a-iv) $\langle \nabla f(x) - \nabla f(y), x - y\rangle \leqslant L\langle \nabla d(x) - \nabla d(y), x - y\rangle$ *для всех* $x, y \in int\, Q$.

Следующие условия эквивалентны:

(b-i) f *— сильно выпуклая относительно d,*

(b-ii) $f - \mu d$ *— выпуклая функция на Q,*

(b-iii) Для дважды дифференцируемых f и d верно $\nabla^2 f(x) \succeq \mu\nabla^2 d(x)$ *для любого* $x \in int\, Q$,

(b-iv) $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geqslant \mu\langle \nabla d(x) - \nabla d(y), x - y\rangle$ *для всех* $x, y \in int\, Q$.

Для удобства применим обозначение $f \preceq d$, если $d - f$ есть выпуклая функция. Это также означает, что f 1-гладкая относительно d ввиду предложения 1.2.3. Аналогично $f \succeq d$ означает, что $f - d$ —

выпуклая функция, и поэтому f 1-сильно выпуклая относительно d (в случае, когда f и d дважды дифференцируемы, соотношение «$\cdot \succeq \cdot$» для двух функций согласуется с частичным порядком Ловнера по гессианам этих двух функций согласно предложению 1.2.3). Тогда условие L-гладкости f относительно d, эквивалентно $f \preceq Ld$; аналогично, условие μ-сильной выпуклости f относительно d эквивалентно $f \succeq \mu d$. Кроме того, свойства относительной гладкости и относительной сильной выпуклости транзитивны: из $f \preceq g$ и $g \preceq d$ следует, что $f \preceq d$.

Известно также следующее утверждение [124] о свойствах суммы и линейных преобразованиях относительно гладких и/или относительно сильно выпуклых функций.

Предложение 1.2.4. *1. Если $f_1 \preceq L_1 d_1$ и $f_2 \preceq L_2 d_2$, то для всяких $\alpha, \beta \geqslant 0$ верно, что*

$$f := \alpha f_1 + \beta f_2 \preceq d := \alpha L_1 d_1 + \beta L_2 d_2.$$

2. Если $f_1 \succeq \mu_1 d_1$ и $f_2 \succeq \mu_2 d_2$, то для всяких $\alpha, \beta \geqslant 0$ верно, что

$$f := \alpha f_1 + \beta f_2 \succeq d := \alpha \mu_1 d_1 + \beta \mu_2 d_2.$$

3. Если $f \preceq d$ и A — линейный оператор в пространстве соответствующей размерности, то $d_f(x) := f(Ax) \preceq d_d(x) := d(Ax)$.

4. Если $f \succeq d$ и A — линейный оператор в пространстве соответствующей размерности, то $d_f(x) := f(Ax) \succeq d_d(x) := d(Ax)$.

В заключение данного пункта упомянем несколько известных работ по оптимизации с идеологией, аналогичной [67] и [124]. В работе [183] имеется единое доказательство для зеркального спуска и проксимального метода при аналогичном предположении об относительной гладкости, в [159] анализируется зеркальный спуск для задач оптимизации в банаховых пространствах при похожих на относительную гладкость предположениях. Более подробное сравнение указанных подходов имеется в [124]. Недавно Ханзели и Рихтарик [104] разработали стохастические алгоритмы для относительно гладких оптимизационных задач.

1.3 Некоторые примеры относительно гладких (сильно выпуклых) задач оптимизации

Приведём несколько примеров [124] оптимизационных задач (1.9), для которых можно построить функцию d, для которой f L-гладкая относительно d и подзадача (1.13) эффективно разрешима.

1.3.1 Задачи оптимизации с полиномиальной зависимостью $\|\nabla^2 f(x)\|$ от $\|x\|_2$

Предположим, что f есть дважды дифференцируемая выпуклая функция на $Q := \mathbb{R}^n$, и $\|\nabla^2 f(x)\|$ — операторная норма от $\nabla^2 f(x)$ относительно ℓ_2-нормы на $\mathbb{R}^n$. Предположим, что $\|\nabla^2 f(x)\| \leqslant p_r(\|x\|_2)$, где $p_r(\alpha) = \sum\limits_{i=0}^{r} a_i \alpha^i$ — многочлен степени r от α. Пусть

$$d(x) := \frac{1}{r+2}\|x\|_2^{r+2} + \frac{1}{2}\|x\|_2^2. \tag{1.16}$$

Следующее предложение утверждает, что f является L-гладкой относительно d, причём значение L нетрудно оценить. Это означает, что независимо от того, насколько быстро растет гессиан f при $\|x\|_2 \to \infty$, f может быть гладкой относительно прокс-функции d (1.16), даже если теряется свойство липшицевости градиента ∇f.

Предложение 1.3.1. *Предположим, что f дважды дифференцируема и верно неравенство $\|\nabla^2 f(x)\| \leqslant p_r(\|x\|_2)$, где $p_r(\alpha)$ — полином степени r от α. Пусть постоянная L такова, что $p_r(\alpha) \leqslant L(1+\alpha^r)$ для всякого $\alpha \geq 0$. Тогда функция f L-гладкая относительно прокс-функции* (1.16).

Учитывая свойство аддитивности в предложении 1.2.3, а также предложение 1.3.1 можно сделать вывод, что практически каждая дважды дифференцируемая выпуклая функция на $\mathbb{R}^n$ является L-гладкой относительно некоторой простой прокс-функции от $\|x\|_2$ полиномиального вида.

Замечание 1.3.2. Предположим, что $p_r(\alpha) = \sum\limits_{i=0}^{r} a_i \alpha^i$. В предложении 1.3.1 наиболее естественный способ оценить L связан с использова-

нием $L = \sum\limits_{i=0}^{r} |a_i|$. Тогда

$$p_r(\alpha) \leqslant \begin{cases} \sum\limits_{i=0}^{r} |a_i|, & \text{для } 0 \leqslant \alpha \leq 1; \\ \sum\limits_{i=0}^{r} |a_i|\alpha^r, & \text{для } \alpha \geq 1. \end{cases}$$

Поэтому $p_r(\alpha) \leqslant \max\{L, L\alpha^r\} \leqslant L(1+\alpha^r)$ для $\alpha \geq 0$.

Важный момент — возможность эффективного решения подзадачи (1.13). Опишем подход к подзадачам вида (1.13) для рассматриваемого класса оптимизационных задач. Ясно, что задача (1.13) может быть записана следующим образом:

$$\min_{x \in \mathbb{R}^n} \langle c, x \rangle + \frac{1}{r+2}\|x\|_2^{r+2} + \frac{1}{2}\|x\|^2,$$

и тогда условие оптимальности первого порядка имеет вид:

$$c + (1 + \|x\|_2^r)x = 0,$$

при этом $x = -\theta c$ для некоторого $\theta \geq 0$. Остаётся лишь определить значение неотрицательного параметра θ. Если $c = 0$, то $x = 0$ удовлетворяет условиям оптимальности. При $c \neq 0$ θ удовлетворяет уравнению первой степени:

$$1 - \theta - \|c\|_2^r \cdot \theta^{r+1} = 0.$$

При $r = 1, 2, 3$ корень такого уравнения возможно найти явно, а при $r > 3$ необходимо уже применять численные методы.

Приведем также более конкретные примеры [124]. Пусть $f(x) := \frac{1}{4}\|Ax - b\|_4^4 + \frac{1}{2}\|Cx - d\|_2^2$ для соответствующих квадратных матриц A и C, а также векторов b и d. Тогда $\nabla^2 f(x) = 3A^T D^2(x) A + C^T C$, где $D(x) = \mathrm{Diag}(Ax - b)$. Покажем, что $f(x)$ L-гладкая относительно прокс-функции

$$d(x) := \frac{1}{4}\|x\|_2^4 + \frac{1}{2}\|x\|_2^2$$

на $Q = \mathbb{R}^n$ для $L = 3\|A\|^4 + 6\|A\|^3\|b\|_2 + 3\|A\|^2\|b\|_2^2 + \|C\|^2$.

Теперь приведём пример относительно гладкой и сильно выпуклой функции, которая не удовлетворяет условию Липшица. Пусть

$$f(x) := \frac{1}{4}\|Ex\|_2^4 + \frac{1}{4}\|Ax - b\|_4^4 + \frac{1}{2}\|Cx - d\|_2^2,$$

где A, C и E — квадратные матрицы, а b и d — векторы. Обозначим через σ_E и σ_C наименьшие собственные значения E и C соответственно, и предположим, что $\sigma_E > 0$ и $\sigma_C > 0$. Тогда

$$\nabla^2 f(x) = \|Ex\|_2^2 E^T E + 2E^T E x x^T E^T E + 3A^T D^2(x) A + C^T C,$$

где $D(x) = \mathrm{Diag}(Ax - b)$. В [124] показано, что $f(x)$ — L-гладкая и μ-сильно выпуклая относительно прокс-функции

$$d(x) := \frac{1}{4}\|x\|_2^4 + \frac{1}{2}\|x\|_2^2$$

на $Q = \mathbb{R}^n$ для $L = 3\|E\|^4 + 3\|A\|^4 + 6\|A\|^3\|b\|_2 + 3\|A\|^2\|b\|_2^2 + \|C\|^2$ и $\mu = \min\left\{\frac{\sigma_E^4}{3}, \sigma_C^2\right\}$.

1.3.2 Имплементируемость тензорных методов третьего порядка К указанному в предыдущем пункте типу задач примыкает следующий, на наш взгляд, класс важных приложений относительной гладкости и относительной сильной выпуклости к методам выпуклой оптимизации высоких порядков. Так, в последние годы в работах Ю. Е. Нестерова, А. В. Гасникова и их коллег активно развиваются так называемые тензорные методы для задач выпуклой оптимизации (см. [155], [156], [19], а также имеющиеся в этих работах ссылки). Речь идёт о задаче минимизации выпуклого функционала $F : \mathrm{R}^n \to \mathbb{R}$, для которого при некотором $r \geqslant 2$

$$\left\|\nabla^r F(y) - \nabla^r F(x)\right\|_2 \leqslant M_r \|y - x\|_2, \quad x, y \in \mathbb{R}^n, \quad M_r \leqslant \infty.$$

Условие $x, y \in \mathrm{R}^n$ можно заменить условием $x, y \in \left\{z \in \mathrm{R}^n : \ F(z) \leqslant F\left(x^0\right)\right\}$, где x^0 — точка старта. Заметим, что $\nabla^r F(y)$ — тензор ранга r. В частности,

$$\nabla^2 F(x) = \left\{\partial \nabla F(x)/\partial x_j\right\}_{j=1}^n = \left\|\partial^2 F(x)/\partial x_i \partial x_j\right\|_{i,j=1}^n$$

есть матрица Гессе дважды гладкой функции $F(x)$. Аналогично можно определить

$$\nabla^{r+1} F(x) = \left\{\partial \nabla^r F(x)/\partial x_j\right\}_{j=1}^n.$$

Поясним, что понимается под 2-нормой от тензора. Ограничимся случаем $r = 2$, тогда

$$\nabla^2 F(x) = \left\|\partial^2 F(x)/\partial x_i \partial x_j\right\|_{i,j=1}^n,$$

$$\begin{gathered}\left\|\nabla^2 F(y) - \nabla^2 F(x)\right\|_2 = \\ = \sup_{\|h_1\|_2 \leqslant 1} \sup_{\|h_2\|_2 \leqslant 1} \left\langle \left(\nabla^2 F(y) - \nabla^2 F(x)\right)[h_1], h_2 \right\rangle = \\ = \sup_{\|x_1\|_2 \leqslant 1} \sup_{\|x_2\|_2 \leqslant 1} \left\langle \left(\nabla^2 F(y) - \nabla^2 F(x)\right) h_1, h_2 \right\rangle = \\ = \max\left\{\lambda_{\max}\left(\nabla^2 F(y) - \nabla^2 F(x)\right), \left|\lambda_{\min}\left(\nabla^2 F(y) - \nabla^2 F(x)\right)\right|\right\}.\end{gathered}$$

Использование информации высоких порядков позволяет улучшить оценки необходимого количества итераций для достижения приемлемого качества решения задачи. Точнее говоря, для класса методов, у которых на каждой итерации разрешается не более чем $O(1)$ раз обращаться к оракулу (подпрограмме) за значениями $\nabla^r F(x)$, $r \leqslant p$, $p \geq 2$, оценка числа итераций, необходимых для достижения точности ε (по функции), будет иметь вид

$$O\left(\min\left\{n \ln\left(\frac{\Delta F}{\varepsilon}\right), \frac{M_0^2 R^2}{\varepsilon^2}, \left(\frac{M_1 R^2}{\varepsilon}\right)^{1/2}, \left(\frac{M_1 R^3}{\varepsilon}\right)^{2/7}, \ldots, \left(\frac{M_p R^{p+1}}{\varepsilon}\right)^{2/(3p+1)}\right\}\right) \tag{1.17}$$

Указанная оценка достижима для недавно предложенного в [19] метода Монтейро–Свайтера–Нестерова. Этот метод использует только $\nabla^r F(x)$, $r \leqslant p$ и сходится согласно оценке (1.17) (для случая $p = 2$ такой метод предложен в работе [129]). Заметим, что в общем случае оценка (1.17) не может быть улучшена, даже если дополнительно известно, что $M_{p+1} < \infty$, $M_{p+2} < \infty, \ldots$.

Стоит заметить, что улучшение оценки необходимого количества итераций компенсируется большими затратами памяти при реализации шага таких тензорных методов. Поэтому практическое использование тензорных методов в пространствах размерности больше 10^3 в настоящее время представляется проблематичным.

В то же время оказывается, что к решению вспомогательных подзадач, которые возникают на итерациях тензорных методов, можно подходить с использованием концепций относительной гладкости и относительной сильной выпуклости. Например, рассмотрим шаг тензорного метода третьего порядка (при $p = 3$)

$$T_{3,3M_3}^{F_{L,x}}(x) =$$

$$= \arg\min_{y\in\mathbb{R}^n}\left\{\sum_{r=0}^{3}\frac{1}{r!}[\nabla_z^r F_{L,x}(z)]_{z=x}\underbrace{[y-x,\ldots,y-x]}_{r}+\frac{3M_3}{4!}\|y-x\|_2^4\right\}=$$

$$= \arg\min_{y\in\mathbb{R}^n}\left\{F_{L,x}(x)+\langle\nabla F_{L,x}(x),y-x\rangle+\frac{1}{2}\langle\nabla^2 F_{L,x}(x)(y-x),y-x\rangle+\right.$$

$$\left.+\frac{1}{6}\nabla^3 F_{L,x}(x)[y-x,y-x,y-x]+\frac{M_3}{8}\|y-x\|_2^4\right\}.$$

Оказывается, что на базе только лишь информации первого и второго порядков ($\nabla F_{L,x}(x)$, $\nabla^2 F_{L,x}(x)$) вполне возможно предложить достаточно эффективный способ приближенного нахождения $T_{3,3M_3}^{F_{L,x}}(x)$. Прежде всего заметим, что нет необходимости вычислять тензор $\nabla^3 F_{L,x}(x)$, потому что в описываемом далее подходе используется только лишь его приближение

$$\nabla^3 F_{L,x}(x)[y-x,y-x]\approx\frac{\nabla^2 F_{L,x}(x+\tau(y-x))(y-x)-\nabla^2 F_{L,x}(x)(y-x)}{\tau}.$$

Далее будет полезно условие относительной сильной выпуклости, о котором шла речь немного ранее. При указанных условиях можно рассмотреть градиентный спуск для вспомогательных подзадач минимизации необходимых функционалов f (с неточным проектированием согласно (0.11))

$$x^{k+1}=\arg\min_{x\in\mathbb{R}^n}^{\widetilde{\delta}}\{\langle\nabla f(x^k),x-x^k\rangle+LV(x,x^k)\}.$$

Он сходится со следующей оценкой

$$\min_{k=0,\ldots,N} f(x^k)-f(x_*)\leqslant(L+\mu)V(x_*,x^0)\left\{\frac{1}{N},\left(\frac{L-\mu}{L+\mu}\right)^N\right\}+\widetilde{\delta},$$

где под $\widetilde{\delta}$ понимается точность решения вспомогательной подзадачи в смысле (0.11). Для рассматриваемой оптимизационной задачи (для простоты обозначений будем считать, что необходимо $T_{3,3M_3}^{F_{L,x}}(0)$) в [154] была предложена следующая прокс-функция

$$d(y)=\frac{1}{2}\langle\nabla^2 f(0)y,y\rangle+\frac{M_3}{8}\|y\|_2^4.$$

В случае такого выбора прокс-функции в [156] было установлено, что

$$\mu=1-\frac{1}{\sqrt{2}},\quad L=1+\frac{1}{\sqrt{2}}.$$

Поэтому решение задачи отыскания $T^{F_{L,x}}_{3,3M_3}(0)$ по сложности эквивалентно (с точностью до логарифмического множителя от желаемой точности нахождения $T^{F_{L,x}}_{3,3M_3}(0)$) задаче вида

$$\langle b, y\rangle + \frac{1}{2}\langle Ay, y\rangle + \lambda\|y\|_2^4 \to \min_{y\in\mathbb{R}^n},$$

что, в свою очередь, эквивалентно задаче

$$\langle b, y\rangle + \frac{1}{2}\langle Ay, y\rangle \to \min_{\|y\|_2^4 \leqslant C(\lambda)},$$

или

$$\langle b, y\rangle + \frac{1}{2}\langle Ay, y\rangle \to \min_{\|y\|_2^2 \leqslant \sqrt{C(\lambda)}}.$$

В итоге рассматриваемая задача эквивалентна

$$\langle b, y\rangle + \frac{1}{2}\langle Ay, y\rangle + \widetilde{\lambda}(\lambda)\|y\|_2^2 \to \min_{y\in\mathbb{R}^n},$$

где $\widetilde{\lambda}(\lambda)$ можно найти за логарифмическое (линейное) время от желаемой точности. Последняя задача имеет такую же сложность (не больше), как и итерация метода Ньютона. Таким образом, с точностью до квадрата логарифмического по желаемой точности множителя задача поиска $T^{F_{L,x}}_{3,3M_3}(x)$ по сложности эквивалентна выполнению итерации метода Ньютона. Более того, как и в методе Ньютона для функции используется информация не выше второго порядка $\nabla F_{L,x}(x),\ \nabla^2 F_{L,x}(x)$. Другими словами, с описанным здесь внутренним методом второго порядка получается, что тензорный метод третьего порядка в такой реализации становится, по сути, методом второго порядка. Полученные в результате такого подхода методы были названы Ю. Е. Нестеровым супербыстрыми тензорными методами второго порядка [155], [156].

$$\frac{\mu}{\mu+2\gamma}\nabla^2 d(x) \preccurlyeq \nabla^2 f(x) \preccurlyeq \nabla^2 d(x)$$

В таком случае предложенный метод будет сходиться с оценкой вида

$$\min_{k=0,\ldots,N} f(x^k) - f(x_*) \leqslant \frac{\mu}{\mu+2\gamma} V(x_*, x^0)\left\{\frac{1}{N}, \left(\frac{\gamma}{\mu+\gamma}\right)^N\right\} + \widetilde{\delta}.$$

1.3.3 Задача D-optimal design [124]. Для матрицы $H \in \mathbb{R}^{m\times n}$ ранга m, где $n \geqslant m+1$, задача D-optimal design имеет вид:

$$f^* = \min_x f(x) := -\ln\det\left(HXH^T\right) \quad \text{при } \langle e, x\rangle = 1 \quad x \geq 0, \tag{1.18}$$

где $X := \mathrm{Diag}(x)$ — диагональная матрица. В вычислительной геометрии задача D-optimal design возникает как лагранжева двойственная задача к хорошо известной задаче о покрывающем эллипсоиде наименьшего объема. Задача (1.18) применима в вычислительной статистике и интеллектуальном анализе данных. По данной задаче накопилось уже довольно немало работ, основанных на применении методов внутренней точки. В работе [124] показана применимость градиентного метода к задаче (1.18). Ясно, что (1.18) — частный случай (1.9) при

$$Q = \Delta_n := \{x \in \mathbb{R}^n : \langle e, x\rangle = 1, \quad x \geq 0\}.$$

Для f из (1.18) выберем прокс-функцию d в виде логарифмического барьера

$$d(x) := -\sum_{j=1}^n \ln(x_j)$$

на положительном ортанте $\mathbb{R}^n_{++}$. Следующее утверждение показывает, что целевой функционал f является 1-гладким относительно d [124].

Предложение 1.3.3. *Предположим, что $f(x) = -\ln\det\left(HXH^T\right)$, где $X = Diag(x)$. Тогда f 1-гладкая относительно $d(x) = -\sum_{j=1}^n \ln(x_j)$ на $\mathbb{R}^n_{++}$.*

Покажем, как возможно решить подзадачу (1.13) на Q при заданной прокс-функции d. Задача (1.13) может быть записана следующим образом:

$$\min_{x\in\Delta_n} \langle c, x\rangle - \sum_{j=1}^n \ln(x_j),$$

и условия оптимальности первого порядка имеют вид:

$$x > 0,\ \langle e, x\rangle = 1 \quad \text{и} \quad c - X^{-1}e = -\theta e$$

для некоторого скалярного множителя θ. Очевидно, что $x_j = 1/(c_j + \theta)$ при $j = 1, \ldots, n$ и остаётся лишь определить значение скаляра θ.

Обратим внимание, что θ должно удовлетворять условию

$$d(\theta) := \sum_{j=1}^{n} \frac{1}{c_j + \theta} - 1 = 0$$

для некоторого θ из интервала $\mathcal{F} := (-\min_j\{c_j\}, \infty)$. Ясно, что d строго уменьшается на $\mathcal{F}$ и $d(\theta) \to +\infty$ при $\theta \searrow -\min_j\{c_j\}$ и $d(\theta) \to -1$ при $\theta \to \infty$, благодаря чему (1.13) имеет единственное решение на интервале $\mathcal{F}$. Кроме того, возможно использовать метод Ньютона (или любой другой подходящий численный метод поиска решения) для эффективного вычисления решения (1.13) на интервале $\mathcal{F}$.

1.3.4 Статистический предобуславливатель для задач минимизации суммы выпуклых функций

Рассмотрим задачу минимизации эмпирического риска вида [107] (см. также [16])

$$\Phi(x) := F(x) + \psi(x) \to min, \quad x \in \mathbb{R}^d, \tag{1.19}$$

где F — эмпирический риск для набора данных $\{z_1, \ldots, z_N\}$:

$$F(x) = \frac{1}{N} \sum_{i=1}^{N} \ell(x, z_i), \tag{1.20}$$

и ψ — выпуклый регуляризирующий функционал. Гладкие регуляризаторы типа евклидовых норм $(\lambda/2)\|x\|^2$ включаются в отдельные функции потерь $\ell(x, z_i)$, а под ψ будем понимать преимущественно негладкие регуляризаторы (например, ℓ_1-норма или индикаторная функция набора ограничений).

В современных приложениях машинного обучения набор данных часто очень большой и его необходимо хранить на нескольких машинах. Для простоты изложения мы предполагаем $N = mn$, где m — количество машин, а n — количество выборок, хранящихся на каждой машине. Пусть $\mathcal{D}_j = \{z_1^{(j)}, \ldots, z_n^{(j)}\}$ обозначает набор данных на машине j и определяет локальный эмпирический риск

$$f_j(x) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(x, z_i^{(j)}\big), \quad j = 1, \ldots, m. \tag{1.21}$$

Общий эмпирический риск (1.20) может быть найден следующим образом

$$F(x) = \frac{1}{m}\sum_{j=1}^{m} f_j(x) = \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n} \ell\big(x, z_i^{(j)}\big).$$

Пусть функционал F L_F-гладкий и σ_F-сильно выпуклый на $\operatorname{Dom}\psi$, то есть

$$\sigma_F I_d \preceq \nabla^2 F(x) \preceq L_F I_d, \quad \forall x \in \operatorname{Dom}\psi,$$

где I_d — единичная матрица размера $d \times d$, $\kappa_F = L_F/\sigma_F$ — стандартное число обусловленности F.

Стандартная для задач распределённой оптимизации предположение заключается в том, что m машин (рабочих) вычисляют градиенты параллельно (независимо), а централизованный сервер по информации от них обновляет переменную x. В частности, на каждой итерации $t = 0, 1, 2, \ldots$,

- сервер передает x_t на все m машин;
- каждая машина j вычисляет градиент $\nabla f_j(x_t)$ и отправляет его обратно на сервер;
- сервер формирует $\nabla F(x_t) = \frac{1}{m}\sum_{j=1}^{m} \nabla f_j(x_t)$ и использует его для вычисления следующей итерации x_{t+1}.

Стандартный подход к задаче (1.19) заключается в реализации проксимального градиентного метода на сервере:

$$x_{t+1} = \arg\min_{x\in\mathbb{R}^d}\left\{\langle \nabla F(x_t), x\rangle + \psi(x) + \frac{1}{2\eta_t}\|x - x_t\|_2^2\right\},$$

где $\|\cdot\|_2$ — евклидова норма, а $\eta_t > 0$ — размер шага. Выбор шага $\eta_t = 1/L_F$ позволяет гарантировать линейную скорости сходимости:

$$\Phi(x_t) - \Phi(x_*) \leqslant \left(1 - \kappa_F^{-1}\right)^t \tfrac{L_F}{2}\|x_* - x_0\|_2^2,$$

где $x_* = \arg\min \Phi(x)$. Другими словами, для достижения качества решения $\Phi(x_t) - \Phi(x_*) \leqslant \varepsilon$ необходимо $O(\kappa_F \log(1/\varepsilon))$ итераций, что также задаёт количество раундов связи между рабочим и сервером. Если мы используем ускоренные методы проксимального градиента на сервере, то оценку на необходимое количество итераций можно уменьшить до $O(\sqrt{\kappa_F}\log(1/\varepsilon))$.

Как известно, для минимизации $F(x) = (1/m)\sum_{j=1}^m f_j(x)$ с использованием методов первого порядка оценку сложности (необходимое количество итераций) $O(\sqrt{\kappa_F}\log(1/\varepsilon))$ не возможно улучшить. Однако для конкретной постановки распределенной минимизации эмпирического риска (ERM) дополнительная структура с конечной суммой каждого f_j в (1.21) позволяет улучшить оценки сложности. Ключевым моментом здесь является то, что если наборы данных $\mathcal{D}_j$ у разных рабочих являются независимыми одинаково распределёнными случайными величинами из одного источника распределения, то локальные эмпирические потери f_j статистически очень похожи друг на друга и на их среднее значение F, особенно когда n достаточно велико.

Статистический предобуславливатель — метод, позволяющий снизить сложность связи на основе указанной идеи. Важным инструментом для статистической предобусловленности методов первого порядка является дивергенция (расхождение) Брегмана и относительная гладкость, которые вводились нами ранее. Рассмотрим задачу [107]

$$f(x) = \frac{1}{m}\sum_{l=1}^m f_l(x) \to \min_{x\in\mathbb{R}^n},$$

где $f_l(x)$ — μ-сильно выпуклые в 2-норме гладкие функции. Будем считать, что m — большое число. Предположим, что есть централизованная архитектура с $r \ll m$ узлами. Первый узел — центральный, то есть связан со всеми остальными. Поместим в первый узел случайно отобранные $\widetilde{m}$ $(\widetilde{m} \ll m)$ слагаемых из суммы. Остальные слагаемые распределим по остальным узлам. Обозначим соответствующую первому узлу нормированную подсумму через $\widetilde{f}(x)$. Рассмотрим градиентный спуск, выполняемый на центральном (первом) узле

$$x^{k+1} = \underset{x\in\mathbb{R}^n}{\overset{\widetilde{\delta}}{\arg\min}} \left\{\left\langle \nabla f(x^k), x - x^k\right\rangle + LV(x, x^k)\right\}, \tag{1.22}$$

где $\widetilde{\delta}$ — точность решения вспомогательных подзадач согласно (0.11), а прокс-функция

$$d(x) = \widetilde{f}(x) + \frac{\gamma}{2}\|x\|_2^2, \quad \gamma > 0.$$

Каждая итерация такого градиентного спуска отвечает коммуникации центрального узла с остальными для нахождения $\nabla f(x^k)$.

Оказывается [107], что если $\|\nabla^2 \widetilde{f}(x) - \nabla^2 f(x)\|_2 \leqslant \gamma$ (здесь под нормой разности гессианов имеется в виду матричная 2-норма, согласованная с векторной 2-нормой), то мы имеем дело с относительно гладкой и сильно выпуклой задачей:

$$\frac{\mu}{\mu + 2\gamma} \nabla^2 d(x) \preccurlyeq \nabla^2 f(x) \preccurlyeq \nabla^2 d(x).$$

В таком случае предложенный метод (1.22) будет сходиться с линейной скоростью (знаменатель прогрессии определяется числом обусловленности).

1.4 Понятия неточного оракула и абстрактной неточной модели целевой функции для задач минимизации функционалов

Далее приводятся необходимые вспомогательные сведения об известном понятии (концепции) неточного оракула О. Деволдера–Ф. Глинера–Ю. Е. Нестерова для оптимизационных задач [90,91]. Говорят, что функция f допускает неточный оракул $(f_\delta(x), g_\delta(x)) \in \mathbb{R} \times E^*$, если выполняется некоторый аналог неравенства (1.8):

$$\begin{aligned} f_\delta(x) + \langle g_\delta(x), y - x \rangle \leqslant f(y) \leqslant \\ \leqslant f_\delta(x) + \langle g_\delta(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 + \delta \ \ \forall x, y \in Q. \end{aligned} \tag{1.23}$$

По сути, (1.23) означает, что $f_\delta(x)$ есть некоторое приближенное значение $f(x)$, а $g_\delta(x)$ — некоторый δ-субградиент f в точке x. Оказывается, что для неускоренного градиентного метода при выполнении условия (1.23) для градиентного метода (с заменой пары $(f, \nabla f)$ на (f_δ, g_δ)) верна такая оценка скорости сходимости:

$$f(\widehat{x}) - f^* \leqslant \frac{C_1}{N} + 2\delta, \tag{1.24}$$

то есть соответствующие погрешностям величины не накапливаются. При этом для быстрого градиентного метода (с заменой пары $(f, \nabla f)$ на (f_δ, g_δ)) выполняется следующая оценка скорости сходимости для некоторой постоянной $C_2 > 0$:

$$f(\widehat{x}) - f^* \leqslant \frac{C_2}{(N+1)^2} + N\delta. \tag{1.25}$$

Сравнение оценок (1.24) и (1.25) указывает на неочевидность преимущества использования ускоренного метода при наличии погрешностей.

Идеология Деволдера–Глинера–Нестерова развита в недавней работе [22], где было предложено обобщение концепции (δ, L)-оракула — (δ, L)-модель целевой функции. Суть подхода в том, что линейная функция $\langle \nabla f(y), x - y \rangle$ в (1.23) заменяется на некоторую абстрактную выпуклую функцию $\psi(x, y)$.

Определение 1.4.1. Будем говорить, что функция f допускает (δ, L)-модель в точке $x \in Q$, и обозначать эту модель $(f_\delta(x), \psi(y, x))$, если для любого $y \in Q$ справедливо неравенство

$$0 \leqslant f(y) - f_\delta(x) - \psi(y, x) \leqslant \frac{L}{2}\|y - x\|^2 + \delta,$$

где $\psi_\delta(x, x) = 0 \ \forall x \in Q$ и $\psi(x, y)$ — выпуклая функция по $x \ \forall y \in Q$.

Концепция из определения 1.4.1 позволяет обосновать сходимость градиентного метода выпуклой минимизации для достаточно широкого класса задач оптимизации [22, 168]. По сути, указанный подход позволяет унифицировать подходы к различным на первый взгляд классам задач оптимизации с проработкой вопросов влияния погрешностей данных на гарантированное качество решения, которого можно достичь в ходе работы метода.

1.4.1 Задачи композитной оптимизации В качестве естественных примеров задач, в которых можно использовать концепцию модели оптимизируемой функции, отметим популярные в анализе данных задачи композитной оптимизации вида

$$f(x) \stackrel{\text{def}}{=} g(x) + h(x) \to \min_{x \in Q}, \tag{1.26}$$

где $g(x)$ — μ-сильно выпуклая и гладкая функция с L-липшицевым градиентом, $h(x)$ — выпуклая функция простой структуры (необязательно гладкая).

К одному из наиболее известных примеров задач вида (1.26) можно отнести так называемую задачу LASSO (Least Absolute Shrinkage and Selection Operator), мотивированную задачами статистики:

$$\frac{1}{2}||Ax - b||_2^2 + \lambda||x||_1 \to \min_{x \in \mathbb{R}^n},$$

где A — матрица размерности $(m \times n)$, $b \in \mathbb{R}^m$, λ — параметр регуляризации, а $||\cdot||_1$ обозначает стандартную l_1-норму и $||\cdot||_2 = \sqrt{\langle x, x\rangle}$. По сути, речь идёт о задаче линейной регрессии по методу наименьших квадратов, но с добавлением ℓ_1-нормы как регуляризатора (именно это слагаемое и рассматривается как композит). Использование такой регуляризации — один из подходов к проблеме неустойчивости стандартной регрессионной модели, связанной с методом наименьших квадратов. Суть проблемы в том, что метод наименьших квадратов, как известно, может быть сильно чувствительным к аномальным данным измерений.

Также в качестве конкретного примера задачи композитной оптимизации можно рассмотреть задачу восстановления матрицы корреспонденций по замерам потоков на линках (ребрах) в большой компьютерной сети (Minimal Mutual Information Model), которая сводится к задаче композитной оптимизации:

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \mu \sum_{k=1}^{n} x_k \ln x_k \to \min_{x \in S_n(1)},$$

где $S_n(1)$ — единичный симплекс в n-мерном пространстве.

Для задач вида (1.26) верно следующее неравенство

$$0 \leqslant f(y) - f(x) - \langle \nabla\varphi(x), y - x\rangle - h(x) + h(y) \leqslant$$

$$\leqslant \frac{L}{2}\|y - x\|^2, \ \ \forall x, y \in Q.$$

В этом случае возможно рассмотреть такую модель

$$\psi(y, x) = \langle \nabla f(x), y - x\rangle + h(y) - h(x), \quad f_\delta(y) = f(y), \ \delta = 0.$$

Это означает, что для данной задачи обычный и ускоренный вариант градиентного метода будут работать без изменений. Если для гладкого случая во вспомогательной задаче стоит функция вида

$$V(x, u_N) + \alpha_{N+1}\langle \nabla\varphi(y_{N+1}), x - y_{N+1}\rangle,$$

то в задаче

$$f(x) := \varphi(x) + h(x) \to \min_{x \in Q}$$

(где $f(x)$ — гладкая выпуклая функция с L-липшицевым градиентом относительно нормы $\|\cdot\|$ и $d(x)$ — выпуклая функция (в общем случае

негладкая)) добавляется слагаемое $h(x)$, и в конечном счете на каждом шаге необходимо решать более сложную задачу

$$V(x, u_N) + \alpha_{N+1} \left(\langle \nabla \varphi(y_{N+1}), x - y_{N+1} \rangle + h(x) - h(y_{N+1}) \right).$$

Если функция h имеет простую структуру (операция проектирования на множестве уровня h не очень затратна), то указанное усложнение вспомогательной задачи может считаться несущественным.

1.4.2 Задача описания модели электоральных процессов Также можно в этом плане отметить недавно предложенную Ю. Е. Нестеровым задачу для модели описания электоральных процессов [144]. В этой модели избиратели (точки данных) выбирают партию (кластер) итеративным способом путем альтернативной минимизации следующей функции

$$f_{\mu_1,\mu_2}(x = (z, p)) = g(x) + \mu_1 \sum_{N=1}^{n} z_N \ln z_N + \frac{\mu_2}{2} \|p\|_2^2 \to \min_{z \in S_n(1), p \in \mathbb{R}^m_+},$$

где $\mathbb{R}^m_+$ — неотрицательный ортант и $S_n(1)$ является стандартным n-мерным симплексом в $\mathbb{R}^n$. Вектор z содержит вероятности, с которыми избиратели выбирают рассматриваемую партию, а вектор p описывает положение партии в пространстве мнений избирателей. Минимизированный потенциал является результатом объединения двух задач оптимизации в одну: избиратели выбирают партию, позиция которой наиболее близка к их личному мнению, а партия корректирует свою позицию, минимизируя дисперсию и стараясь не отходить слишком далеко от своей первоначальной позиции. Ю. Е. Нестеров использовал последовательный процесс выборов, чтобы показать, что при некоторых естественных допущениях процесс сближается и дает кластеризацию точек данных. Это было сделано в [144] для конкретного выбора функции g с ограниченной интерпретируемостью. Как показано далее в разделе 4.2, общая концепция модели целевой функции позволяет применить рассматриваемые в настоящей работе методы градиентного типа в случае не обязательно выпуклых функционалов g.

1.4.3 Некоторые другие примеры [22]

Пример 1.4.2. (Суперпозиция функций) Рассмотрим следующую задачу

$$f(x) := f(f_1(x), \ldots, f_m(x)) \to \min_{x \in Q},$$

где $f_N(x)$ — гладкая выпуклая функция с L_N-липшицевым градиентом в норме $\|\cdot\|$. $f(x)$ — M-липшицева выпуклая функция относительно L_1-нормы, неубывающая по каждому из своих аргументов. Как следствие [94], f — также выпуклая функция, и верно следующее неравенство

$$0 \leqslant f(y) - f(f_1(x) + \langle \nabla f_1(x), y - x \rangle, \ldots, f_m(x) + \langle \nabla f_m(x), y - x \rangle) \leqslant$$
$$\leqslant M \frac{\sum_{i=1}^m L_i}{2} \|y - x\|^2, \, \forall x, y \in Q.$$

И верно

$$0 \leqslant f(y) - f(x) - f(f_1(x) +$$
$$+ \langle \nabla f_1(x), y - x \rangle, \ldots, f_m(x) + \langle \nabla f_m(x), y - x \rangle) + f(x) \leqslant$$
$$\leqslant M \frac{\sum_{i=1}^m L_i}{2} \|y - x\|^2, \, \forall x, y \in Q.$$

Можно положить

$$\psi(y, x) = f(f_1(x) + \langle \nabla f_1(x), y - x \rangle, \ldots, f_m(x) + \langle \nabla f_m(x), y - x \rangle) - f(x),$$

$$f_\delta(x) = f(x) \text{ и } \delta = 0.$$

Как и в задаче (1.4.1), оценки на скорость сходимости сохраняются, но при этом вспомогательные задачи могут сильно усложниться. Данная задача может включать обширное количество частных случаев [93, 94]: гладкая оптимизация, негладкая оптимизация, минимаксная задача [63], композитная оптимизация, задача с регуляризацией.

Рассмотрим без подробного объяснения ещё несколько примеров постановок задач, в которых может быть актуальной рассматриваемая нами концепция модели функции.

Пример 1.4.3. (Минмин задача) Рассматривается следующая задача:

$$f(x) \stackrel{\text{def}}{=} \min_{y \in Q} f(y, x) \to \min_{x \in \mathbb{R}^n}.$$

Пусть $f(y,x)$ — гладкая и $\forall y, y' \in Q$, $\forall x, x' \in \mathbb{R}^n$ выполнено:

$$\|\nabla f(y',x') - \nabla f(y,x)\|_2 \leqslant L \|(y',x') - (y,x)\|_2 .$$

Тогда (из [96]) если можно найти такую $\widetilde{y}_\delta(x) \in Q$, что

$$\langle \nabla_y f(\widetilde{y}_\delta(x), x), y - \widetilde{y}_\delta(x) \rangle \geqslant -\delta, \ \ \forall y \in Q,$$

то

$$f(\widetilde{y}_\delta(x), x) - f(x) \leqslant \delta, \|\nabla f(x') - \nabla f(x)\|_2 \leqslant L \|x' - x\|_2 ,$$

и

$$(f_\delta(x) = f(\widetilde{y}_\delta(x), x) - 2\delta, \psi(z,x) = \langle \nabla_y f(\widetilde{y}_\delta(x), x), z - x \rangle)$$

будет $(6\delta, 2L)$-моделью для функции $f(x)$ в точке x. Таким образом, мы получаем $(6\delta, 2L)$-модель, которая может быть использована для решения (1.4.3).

Пример 1.4.4. (Седловая задача) Рассматривается следующая задача нахождения седловой точки [97]

$$f(x) \stackrel{\text{def}}{=} \max_{y \in Q} [\langle x, b - Ay \rangle - d(y)] \to \min_{x \in \mathbb{R}^n},$$

где $d(y)$ — μ-сильно выпуклая относительно p-нормы ($1 \leqslant p \leq 2$). Тогда (из [79]) $f(x)$ — гладкая функция с константой Липшица градиента в 2-норме

$$L = \frac{1}{\mu} \max_{\|y\|_p \leq 1} \|Ay\|_2^2 .$$

Если $y_\delta(x)$ — решение вспомогательной задачи максимизации с точностью по функции δ, то

$$(f_\delta(x) = \langle x, b - Ay_\delta(x) \rangle - d(y_\delta(x)), \psi(z,x) = \langle b - Ay_\delta(x), z - x \rangle)$$

будет $(\delta, 2L)$-моделью для функции $f(x)$ в точке x.

Пример 1.4.5. (Прокс-метод с неточным решением задачи минимизации на итерации) Рассматривается следующая функция

$$f(x) := \min_{y \in Q} \underbrace{\left\{ d(y) + \frac{L}{2} \|y - x\|_2^2 \right\}}_{\Lambda(x,y)} .$$

Пусть $d(y)$ — выпуклая функция и

$$\max_{y \in Q} \left\{ \Lambda(x, y(x)) - \Lambda(x, y) + \frac{L}{2} \|y - y(x)\|_2^2 \right\} \leqslant \delta$$

Тогда верно, что

$$\left(f_\delta(x) = d(y(x)) + \frac{L}{2} \|y(x) - x\|_2^2 - \delta, \psi(z, x) = \langle L(x - y(x)), z - x \rangle \right)$$

есть (δ, L)-модель для функции f в точке x.

Для достаточно гладких бесконечномерных задач [101] на гильбертовых пространствах хорошо известно, как возможно вполне естественно применять (адаптивные и неадаптивные) градиентные методы с использованием неточного оракула Деволдера–Глинера–Нестерова.

1.5 Зеркальный спуск для задач выпуклой оптимизации с функционалами, которые могут не удовлетворять условию Липшица. Условие относительной липшицевости.

Как правило, при анализе сходимости методов первого порядка для задач негладкой (недифференцируемой) выпуклой оптимизации предполагается, что целевая функция удовлетворяет условию Липшица. Однако не всегда такое условие верно. Например, рассмотрим хорошо известную задачу, которая связана с методом опорных векторов (SVM) для бинарной классификации в машинном обучении, оптимизационная формулировка которой имеет вид:

$$\min_x f(x) := \frac{1}{2} \sum_{i=1}^{n} \max \left\{ 0, 1 - y_i x^T \omega_i \right\} + \frac{\lambda}{2} \|x\|_2^2, \tag{1.27}$$

где ω_i — вектор входных признаков выборки i, а $y_i \in \{-1, 1\}$ — метка выборки i. Ясно, что f недифференцируем и не удовлетворяет условию Липшица (ввиду наличия параметра регуляризации, связанного с

2-нормой). Таким образом, для такой задачи нет возможности напрямую использовать типичные субградиентные схемы и соответствующие оценки скорости сходимости.

Другим примером является проблема вычисления общей точки пересечения $x \in R^m$ n эллипсоидов. Оптимизационная постановка задачи имеет следующий вид:

$$f^* = \min_x f(x) := \max_{0 \leqslant i \leqslant n} \left\{ \frac{1}{2} x^T A_i x + b_i^T x + c_i \right\}, \tag{1.28}$$

$$IEP: \ f^* = \min_x f(x) := \max_{0 \leqslant i \leqslant n} \left\{ \frac{1}{2} x^T A_i x + b_i^T x + c_i \right\},$$

где i-ый эллипсоид $Q_i = \left\{ x \in \mathbb{R}^m : \frac{1}{2} x^T A_i x + b_i x + c_i \leqslant 0 \right\}$ и $A_i \in R^{m \times m}$ — симметричная положительная полуопределенная матрица, $i = 1, \ldots, n$. Стоит заметить, что целевая функция f является недифференцируемой и нелипшицевой, и поэтому к ней нет возможности применить стандартные для субградиентных методов оценки.

Однако оказывается [125], что вполне возможно обобщить свойство Липшица, заменив его на липшицевость относительно некоторой выпуклой функции d. При этом, как и в упомянутых выше условиях относительной гладкости, не требуется ни сильной, ни даже строгой выпуклости d. Этот подход в некотором смысле представляет из себя развитие для недифференцируемой выпуклой оптимизации подходов упомянутой выше работы [124] для задач дифференцируемой (гладкой) выпуклой оптимизации.

Отметим, что методы зеркального спуска можно применять к задачам типа (1.27) и (1.28) без использования относительной липшицевости. Рассмотрим схему зеркального спуска с переключениями, для которой возможно выписать оценки для целевых функционалов различного уровня гладкости, в том числе и для задач типа (1.27) и (1.28):

$$M_N = \|\nabla f(x^N)\|_*, \ h_N = \frac{\varepsilon}{M_N} \ x^{N+1} = Mirr_{x^N}(h_N \nabla f(x^N)), \tag{1.29}$$

где $Mirr_{x^N}(h_N \nabla f(x^N))$ определяется в (1.6).

Также в задачах с квадратичными целевыми функционалами мы сталкиваемся с ситуацией, когда такой функционал не удовлетворяет обычному свойству Липшица (или константа Липшица достаточно большая), но градиент удовлетворяет условию Липшица. Можно рассмат-

ривать и более широкий класс уже негладких целевых функционалов

$$f(x) = \max_{1 \leqslant i \leqslant m} f_i(x), \tag{1.30}$$

где

$$f_i(x) = \frac{1}{2}\langle A_i x, x\rangle - \langle b_i, x\rangle + \alpha_i,\ i = 1, \ldots, m, \tag{1.31}$$

в случае, когда A_i $(i = 1, \ldots, m)$ — положительно определённые матрицы: $x^T A_i x \geqslant 0\ \forall x \in Q$. Отметим, что функционалы вида (1.30)–(1.31) возникают в задачах проектирования механических конструкций Truss Topology Design со взвешенными балками.

В таком случае важно использовать следующий подход к выводу оценок скорости сходимости для зеркальных спусков с шагом, сопоставимым с нормализованным градиентом целевого функционала. Для всякого ненулевого конечного субградиента $\nabla f(x)$ целевого функционала f введём следующую вспомогательную величину

$$v_f(x, x_*) = \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|_*}, x - x_* \right\rangle, \quad x \in Q,$$

где x_* — искомое решение задачи. Случай $\nabla f(x) = 0$ здесь опускается, поскольку тогда x — автоматически искомая точка. Для получения оценок полезно следующее известное вспомогательное утверждение [39].

Лемма 1.5.1. *Введем следующую функцию*

$$\omega(\tau) = \max_{x \in Q}\{f(x) - f(x_*) : \|x - x_*\| \leqslant \tau\},$$

где τ — положительное число. Тогда для всякого $y \in Q$

$$f(y) - f(x_*) \leqslant \omega(v_f(y, x_*)).$$

Если для найденной в ходе работы метода точки $y = x^N$ известно, что $v_f(y, x_*) \leqslant \varepsilon$, то с использованием предыдущего утверждения можно оценить скорость сходимости алгоритма (1.29) для дифференцируемого целевого функционала f с градиентом, удовлетворяющим условию Липшица:

$$\|\nabla f(x) - \nabla f(y)\|_* \leqslant L\|x - y\| \quad \forall x, y \in Q. \tag{1.32}$$

Учитывая следующее известное неравенство

$$f(x) \leqslant f(x_*) + \|\nabla f(x_*)\|_*\|x - x_*\| + \frac{1}{2}L\|x - x_*\|^2,$$

получаем, что

$$\min_{N\in I} f(x^N) - f(x_*) \leqslant \min_{N\in I}\left\{\|\nabla f(x_*)\|_*\|x^N - x_*\| + \frac{1}{2}L\|x^N - x_*\|^2\right\}.$$

Далее, верна оценка

$$f(x) - f(x_*) \leqslant \varepsilon\cdot\|\nabla f(x_*)\|_* + \frac{1}{2}L\varepsilon^2.$$

Поэтому справедливо

Следствие 1.5.2. *Пусть f дифференцируема на Q, и верно* (1.32)*. Тогда после остановки рассмотренного метода верно*

$$\min_{1\leqslant N\leqslant N} f(x^N) - f(x_*) \leqslant \varepsilon\cdot\|\nabla f(x_*)\|_* + \frac{L\varepsilon^2}{2}.$$

Также можно рассмотреть специальный класс негладких целевых функционалов.

Следствие 1.5.3. *Пусть $f(x) = \max\limits_{i=\overline{1,m}} f_i(x)$, где f_i дифференцируема для всякого $x\in Q$ и $\|\nabla f_i(x) - \nabla f_i(y)\|_* \leqslant L_i\|x-y\| \quad \forall x,y\in Q$. Тогда после остановки алгоритма зеркального спуска верна оценка*

$$\min_{0\leqslant N\leqslant N} f(x^N) - f(x_*) \leqslant \varepsilon\cdot\|\nabla f(x_*)\|_* + \frac{L\varepsilon^2}{2},$$

где $L = \max\limits_{i=\overline{1,m}} L_i$.

Все упомянутые в настоящем пункте результаты положены в основу главы 5, посвящённой методам зеркального спуска для задач выпуклого программирования при разных условиях на уровень гладкости целевого функционала и функционалов ограничений.

1.5.1 Задача проектирования механических конструкций (Truss Topology Design)

Задача проектирования механических конструкций состоит в нахождении лучшей механической структуры, способной выдержать внешнюю нагрузку, но при этом обладающей минимальными затратами материала на саму конструкцию. Математическая формулировка выглядит так:

$$\min_{w\in R^N_+}\{\langle\overline{f}, u\rangle : A(w)u = \overline{f},\ \langle e, w\rangle = T\}, \tag{1.33}$$

где $\overline{f}$ — вектор внешних сил, $u \in R^{2n}$ — вектор виртуального смешения n узлов в R^2, w — вектор N балок и T — общий вес конструкции. Матрица соответствия $A(w)$ принимает следующую форму:

$$A(w) = \sum_{i=1}^{N} w_i a_i a_i^T,$$

где $a_i \in R^{2n}$ является вектором, описывающим взаимодействие двух узлов, соединенных дугой. Размер таких векторов небольшой: на 2D-модель приходится не более 4 ненулевых элементов.

Опишем подход [167] к указанной задаче (1.33). Её можно переписать в виде задачи линейного программирования

$$\begin{aligned}
&\min_{u,w}\{\langle \overline{f}, u\rangle : A(w)u = \overline{f},\ w \geqslant 0,\ \langle e, w\rangle = T\} = \\
&= \min_{w}\{\langle \overline{f}, A^{-1}(w)\overline{f}\rangle : w \in \triangle(T) = \{ w \geqslant 0,\ \langle e, w\rangle = T\}\} = \\
&= \min_{w\in\triangle(T)} \max_{u}\{2\langle \overline{f}, u\rangle - \langle A(w)u, u\rangle\} \geqslant \max_{u} \min_{w\in\triangle(T)}\{2\langle \overline{f}, u\rangle - \langle A(w)u, u\rangle\} = \\
&= \max_{u}\{2\langle \overline{f}, u\rangle - T \max_{1\leqslant i\leqslant N}\langle a_i, u\rangle^2\} = \max_{\lambda, y}\{2\lambda\langle \overline{f}, y\rangle - \lambda^2 T \max_{1\leqslant i\leqslant N}\langle a_i, y\rangle^2\} = \\
&= \max_{y} \frac{\langle \overline{f}, y\rangle^2}{T \max\limits_{1\leqslant i\leqslant N}\langle a_i, y\rangle^2} = \frac{1}{T}\left(\max_{y}\{\langle \overline{f}, y\rangle : \max_{1\leqslant i\leqslant N}|\langle a_i, y\rangle| \leqslant 1\}\right)^2.
\end{aligned} \tag{1.34}$$

Стоит заметить, что для неравенства в третьей строке не нужно никаких предположений. Обозначим через y^* решение данной задачи оптимизации. Тогда существуют множители $x^* \in R_+^N$, такие, что

$$\overline{f} = \sum_{i\in J_+} a_i x_i^* - \sum_{i\in J_-} a_i x_i^*, \qquad x_i^* = 0,\ i \notin J_+ \bigcap J_-, \tag{1.35}$$

где $J_+ = \{i : \langle a_i, y^*\rangle = 1\}$ и $J_- = \{i : \langle a_i, y^*\rangle = -1\}$. Перемножив первое уравнение в (1.35) на y^*, получим

$$\langle \overline{f}, y^*\rangle = \langle e, x^*\rangle.$$

Заметим, что первое уравнение в (1.35) может быть записано как

$$\overline{f} = A(x^*)y^*. \tag{1.36}$$

Покажем, как можно по решению двойственной задачи восстановить решение основной задачи. Обозначим

$$w^* = \frac{T}{\langle e, x^* \rangle} \cdot x^*, \qquad u^* = \frac{\langle e, x^* \rangle}{T} \cdot y^*. \tag{1.37}$$

Тогда с учетом (1.36) имеем $\overline{f} = A(w^*)u^*$ и $w^* \in \triangle(T)$. Таким образом, пара в (1.37) возможна для основной задачи. С другой стороны,

$$\langle \overline{f}, u^* \rangle = \langle \overline{f}, \frac{\langle e, x^* \rangle}{T} \cdot y^* \rangle = \frac{1}{T} \cdot \langle e, x^* \rangle \cdot \langle \overline{f}, y^* \rangle = \frac{1}{T} \cdot \langle \overline{f}, y^* \rangle^2.$$

Таким образом, зазор двойственности в цепочке (1.34) равен нулю. Поэтому пара векторов (w^*, u^*), определяемая (1.37), есть решение основной задачи.

Приведенные выше рассуждения (см. [167]) позволяет заменить исходную задачу на (двойственную) задачу линейного программирования:

$$\max_y \{ \langle \overline{f}, y \rangle : \max_{1 \leqslant i \leqslant N} \langle \pm a_i, y \rangle \leqslant 1 \}, \tag{1.38}$$

Предположим, что имеется механическая конструкция *локальной* топологии, в которой каждый узел конструкции связан только с восемью соседними узлами. Это позволяет применить свойство *разреженности* для векторов ba_i $(1 \leqslant i \leqslant m)$. В таком случае вычислительные затраты каждой итерации растут как $O(\log_2 m)$ [153, 167].

В [153] рассмотрен специальный класс задач для разреженных задач оптимизации большой размерности, для которых выгоднее рассматривать сведение к негладкой задаче и использовать субградиентные методы. Согласно [153] для квадратичной функции $f(y) = \frac{1}{2}\langle Ay, y \rangle$ операция вычисления градиента $\nabla f(y) = Ay$ связана с умножением на матрицу A, что может быть затратным даже в случае разреженности A. Однако субградиенты негладкой функции $f(y) = \max\limits_{1 \leqslant i \leqslant m} \langle ba_i, y \rangle$ (см. (1.38) выше) будут разреженными векторами, если все векторы ba_i разрежены. Этот факт основан на следующем наблюдении. Для функции $f(y) = \max\limits_{1 \leqslant i \leqslant m} \langle ba_i, y \rangle$ с разреженной матрицей $A = (ba_1, ba_2, ..., ba_m)$ вектор $\nabla f(y) = ba_i(y)$ — субградиент f в точке y. В таком случае стандартный шаг субградиентного метода $y_+ = y - h \cdot \nabla f(y)$ приводит к изменению только нескольких элементов вектора y, и вектор $z_+ = A^T y_+$ отличается от $z = A^T y$ также только по нескольким координатам. Таким образом, значение функции $f(y_+)$ может быть легко обновлено при

условии, что есть эффективная процедура для пересчета максимума m значений.

Обратим внимание, что целевой функционал в (1.38) линеен и затраты на выполнении итерации сопоставимы с рассмотренными в [153,167] простой субградиентной схемой с переключениями. Поэтому основные наблюдения работ [153, 167] верны и для разработанных в главе 5 алгоритмических схем. В этом случае задача (1.38) рассматривается как задача минимизации линейного функционала $f(y) = \langle \overline{f}, y \rangle$ с негладким функционалом ограничения $g(y) = \max_{1 \leqslant i \leqslant N} \langle \pm a_i, y \rangle$. К такой задаче возможно применить субградиентные схемы с переключениями вида $y^{k+1} = y^k - h_k \nabla f(y^k)$, если значение g в точке y^k приемлемо (например, $g(y^k) \leqslant \varepsilon$), и $y^{k+1} = y^k - h_k \nabla g(y^k)$ — в противном случае.

ГЛАВА 2

Адаптивные и универсальные методы для вариационных неравенств и седловых задач

Введение

Вариационные неравенства (ВН) и седловые задачи часто возникают в самых разных проблемах оптимизации и имеют многочисленные приложения в математической экономике, математическом моделировании транспортных потоков, теории игр и других разделах математики (см., например, [99]). Исследования в области алгоритмических методов решения вариационных неравенств и седловых задач активно продолжаются. Наиболее известным аналогом градиентного метода для ВН является экстраградиентный метод Г. М. Корпелевич [32], в качестве одного из современных вариантов которого можно выделить проксимальный зеркальный метод А. С. Немировского [135].

В этой главе мы введём новое понятие (концепцию) неточной оптимизационной модели для вариационных неравенств и седловых задач, которая по сути есть аналог известного для задач минимизации функционалов неточного (δ, L)-оракула О. Деволдера–Ф. Глинера–Ю. Е. Нестерова. Для задач, которые допускают существование такой модели в произвольной точке области определения (это достаточно широкий класс вариационных неравенств и седловых задач), в настоящей главе предложен аналог проксимального зеркального метода с адаптивным выбором шага и адаптивным правилом остановки, выполнение которого гарантирует достижение желаемого качества решения. При этом обосновано сохранение оптимальных вычислительных гарантий (оценок скорости сходимости) для вариационных неравенств с липшицевыми монотонными операторами, а также — с ограниченными монотонными операторами и на соответствующих классах выпукло-вогнутых седловых задач. Помимо этого предложенный подход позволяет учесть возможность неточного задания оператора вариационного неравенства,

причём обосновано отсутствие накопления этих погрешностей в итоговых оценках скорости сходимости. Важно также, что удалось найти интерполяцию для вариационных неравенств с монотонными гёльдеровыми операторами со специально введённой искусственной неточностью, что позволило предложить *универсальный метод* для вариационных неравенств и седловых задач соответствующего уровня гладкости. Универсальность здесь понимается как возможность адаптивной настройки работы метода на уровень гладкости вариационного неравенства. Напомним, что универсальные градиентные методы некоторое время назад предложены в [152] для задач минимизации выпуклого функционала с гёльдеровым (суб)градиентом и неплохо себя зарекомендовали в различных задачах минимизации функционалов с ослабленным условиями гладкости.

Более конкретно выделим классы задач, которым посвящена настоящая глава работы. Для оператора $G : Q \to \mathbb{R}^n$, заданного на выпуклом компакте $Q \subset \mathbb{R}^n$, под *вариационным неравенством* понимаем неравенство вида

$$\langle G(x_*),\ x_* - x\rangle \leqslant 0. \tag{2.1}$$

При этом предположим, что G удовлетворяет условию Гёльдера:

$$||G(x) - G(y)||_* \leqslant L_\nu ||x - y||^\nu \quad \forall x, y \in Q \tag{2.2}$$

для некоторого фиксированного $\nu \in [0; 1]$.

Отметим, что в (2.1) требуется найти $x_* \in Q$ (это x_* и называется (строгим) решением ВН), для которого

$$\max_{x \in Q} \langle G(x_*),\ x_* - x\rangle \leqslant 0.$$

Для монотонного оператора поля G можно рассматривать также задачу отыскания слабого решения ВН

$$\langle G(x),\ x_* - x\rangle \leqslant 0, \tag{2.3}$$

то есть нахождения $x_* \in Q$, такого, что (2.3) верно при всех $x \in Q$. Получена оценка сложности (достаточного количества итераций для достижения приемлемого качества решения) предложенного универсального метода

$$O\left(\left(\frac{1}{\varepsilon}\right)^{\frac{2}{1+\nu}}\right), \tag{2.4}$$

которая оптимальна в случаях $\nu = 0$ и $\nu = 1$ [36,103,136,182]. При этом показано, что адаптивность предлагаемого метода на практике может приводить к ускорению работы метода по сравнению с оценками (2.4).

Подчеркнем, что указанный результат удаётся достичь за счёт введения искусственной неточности в оптимизационную модель. Это привело к идее предложить общую концепцию неточного оракула (введённую несколько лет назад для задач минимизации функционалов) для вариационных неравенств, которая позволила бы применять оптимальные методы на классе ВН с липшицевыми монотонными операторами для задач с погрешностями (необязательно искусственными). В настоящей главе вводится аналог концепции (δ, L)-*модели* функции для вариационных неравенств и седловых задач. Такой подход позволит обосновать применимость подхода [17] к более широкому классу задач, в частности к *смешанным вариационным неравенствам* [31,69] и *композитным седловым задачам* (здесь стоит отметить популярную в оптимизационном сообществе работу [84]). Далее, будем рассматривать задачу нахождения решения $x_* \in Q$ абстрактной задачи равновесного программирования (неравенство Фань Цзы)

$$\psi(x, x_*) \geqslant 0 \quad \forall x \in Q \tag{2.5}$$

для некоторого выпуклого компакта $Q \subset \mathbb{R}^n$, а также функционала $\psi : Q \times Q \to \mathbb{R}$. Если предположить абстрактную монотонность функционала ψ:

$$\psi(x, y) + \psi(y, x) \leqslant 0 \ \ \forall x, y \in Q, \tag{2.6}$$

то всякое решение (2.5) будет также и решением двойственной задачи равновесного программирования

$$\psi(x_*, x) \leqslant 0 \quad \forall x \in Q. \tag{2.7}$$

В общем случае сделаем предположение о существовании решения x_* задачи (2.5). Приведем пару примеров задания ψ, для которых данное условие заведомо выполнено.

Пример 2.0.1. Если для некоторого оператора $G : R \to \mathbb{R}^n$ положить

$$\psi(x, y) = \langle G(y), x - y \rangle \ \ \forall x, y \in Q, \tag{2.8}$$

то (2.5) и (2.7) будут равносильны соответственно стандартным задачам отыскания строгого и слабого решений вариационных неравенств с оператором G.

Пример 2.0.2. Для некоторого оператора $G : Q \to \mathbb{R}^n$ и выпуклого функционала $h : Q \to \mathbb{R}^n$ простой структуры (см., например, [16]) выбор функционала

$$\psi(x, y) = \langle G(y), x - y\rangle + h(x) - h(y)$$

приводит к *смешанному вариационному неравенству* [31, 69]

$$\langle G(y), y - x\rangle + h(y) - h(x) \leqslant 0,$$

которое в случае монотонности оператора G влечет

$$\langle G(x), y - x\rangle + h(y) - h(x) \leqslant 0.$$

Отметим, что известно немало проксимальных методов для задач нахождения точек равновесия (см., в частности, [6,15] и имеющуюся там библиографию). В частности, в [15] предложен проксимальный метод для задач равновесного программирования в гильбертовых пространствах. Однако, как правило, в этих работах лишь исследуются условия сходимости предлагаемых методов без обоснования оптимальности скорости сходимости, а также критериев остановки рассматриваемых методов, гарантирующих достижение приемлемого качества решения. В настоящей главе, среди прочего, обосновывается применимость к таким задачам адаптивного аналога проксимального зеркального метода А. С. Немировского [135], причём в общности неточного оракула. При этом предлагаемый нами адаптивный критерий остановки за конечное число шагов обеспечивает достижение приемлемого качества приближённого решения с гарантированно оптимальными (с точностью) до умножения на константу оценками скорости сходимости для выделенного класса задач (это следует из того, что эти оценки оптимальны на более узком классе вариационных задач с липшицевым оператором).

2.1 Понятие неточной оптимизационной модели для вариационных неравенств

Введем анонсированное ранее понятие (δ, L)-модели для задач вида (2.5) и (2.7).

Определение 2.1.1. Будем говорить, что функционал ψ_δ есть (δ, L)-модель для задач (2.5) и (2.7) на множестве Q, если $\forall x, y, z \in Q$ верны следующие предположения.

(i) $\psi(x,y) \leqslant \psi_\delta(x,y) + \delta$;

(ii) функционал $\psi_\delta(x,y)$ выпуклый по первой переменной;

(iii) $\psi_\delta(x,x) = 0 \ \ \forall x \in Q$;

(iv) (*абстрактная δ-монотонность*) неравенство $\psi_\delta(x,y) + \psi_\delta(y,x) \leqslant \delta$;

(v) (*обобщенная гладкость*)

$$\psi_\delta(x,y) \leqslant \psi_\delta(x,z) + \psi_\delta(z,y) + LV(x,z) + LV(z,y) + \delta \qquad (2.9)$$

для некоторой фиксированной константы $L > 0$, где $\delta > 0$ — некоторая постоянная величина (оценка погрешности задания ψ, степень отклонения от гладкости).

Отметим, что в случае обычного ВН (2.8) и евклидовой нормы условие (2.9) сводится к неравенству

$$\langle G(z) - G(y), z - x\rangle \leqslant \frac{L}{2}||z-x||^2 + \frac{L}{2}||z-y||^2 + \delta. \qquad (2.10)$$

При $\delta = 0$ неравенство (2.10) легко проверяется, например для оператора $G(x) = \nabla f(x)$, где $f : Q \to \mathbb{R}$ есть некоторый выпуклый субдифференцируемый функционал и $\nabla f(x)$ — произвольный субградиент f в точке x. Заметим, что при $\delta = 0$ похожее на (2.9) условие

$$\psi(x,y) \leqslant \psi(x,z) + \psi(z,y) + a\|y-z\|^2 + b\|z-x\|^2 \quad \forall x,y,z \in Q$$

(a и b — положительные константы) предложено в [127] и использовалось во многих последующих работах (см., например, [15] и имеющуюся там библиографию). Предложенный нами подход позволяет работать с неевклидовой прокс-структурой, а также учитывать неточность δ, что важно для идеологии универсальных методов.

Пример 2.1.2. Условие относительной гладкости для задач оптимизации и вариационных неравенств. Рассмотрим задачу минимизации

$$\min_{x \in Q} f(x), \qquad (2.11)$$

где функция f выпуклая и L-гладкая относительно d, то есть для произвольных $x, y \in Q$

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leqslant LV(x, y).$$

В таком случае (2.11) эквивалентна задаче нахождения x_*:

$$\max_{x \in Q} \psi_\delta(x_*, x) \leqslant 0$$

при $\psi_\delta(x, y) := \langle \nabla f(y), x - y \rangle$. Свойства (i)-(iv) определения 2.1.1, очевидно, выполняются при $\delta = 0$. Проверим, что (v) также выполнено. В самом деле,

$$\begin{aligned}
&\psi_\delta(x, y) - \psi_\delta(x, z) - \psi_\delta(z, y) = \\
&= \langle \nabla f(y), x - y \rangle - \langle \nabla f(z), x - z \rangle - \langle \nabla f(y), z - y \rangle = \\
&= (f(x) - f(z) - \langle \nabla f(z), x - z \rangle) + (f(z) - f(y) - \langle \nabla f(y), z - y \rangle) - \\
&- (f(x) - f(y) - \langle \nabla f(y), x - y \rangle) \leqslant LV(x, z) + LV(z, y),
\end{aligned}$$

где мы использовали относительную L-гладкость и выпуклость f. Этот пример показывает, что предложенный нами подход к понятию неточной оптимизационной модели для ВН применима и в случае относительной гладкости оператора (это аналог известного условия относительной гладкости для задач минимизации функционалов). В этом конкретном случае будем говорить, что оператор G относительно L-гладкий, если

$$\langle G(y) - G(z), x - z \rangle \leqslant LV(x, z) + LV(z, y) \ \ \forall x, y, z \in Q.$$

Пример 2.1.3. Интересный пример задачи совместного использования ресурсов, в которой естественно возникает вариационное неравенство с относительно гладким и монотонным оператором приведён в [65] (пример 2.3). Как показано в [65] (пример 3.5), возникающий в данной задаче оператор вариационного неравенства $\widehat{G}$ удовлетворяет обобщённому условию Липшица вида

$$\|\widehat{G}(y) - \widehat{G}(x)\|_{loc} \leqslant \sqrt{V(x, y)} \quad \forall x, y \in Q$$

для некоторого аналога нормы (локальной нормы) $\|\cdot\|_{loc}$ при специальном выборе дивергенции V. При этом $V(y, x) \geqslant \|y - x\|^2_{loc}$ для произвольных $x, y \in Q$. Это означает, что оператор $\widehat{G}$ относительно $\frac{1}{2}$-гладкий на Q:

$$\langle \widehat{G}(y) - \widehat{G}(z), x - z \rangle \leqslant \sqrt{V(z, y)} \|x - z\|_{loc} \leqslant \frac{1}{2}(V(x, z) + V(z, y) \ \ \forall x, y, z \in Q.$$

Это означает, что разработанные нами методы для вариационных неравенств и седловых задач в настоящей работе применимы и к такой задаче, причём в общности неточной оптимизационной модели (с неточными данными).

Рассмотрим ещё пример, связанный с модельной общностью.

Пример 2.1.4. Предположим, что функционал $f : Q \to \mathbb{R}$ допускает в каждой точке $x \in Q$ (δ, L)-модель $\psi(y, x)$. В таком случае для всяких $x, y, z \in Q$ верны неравенства

$$f(x) \leqslant f(z) + \psi(x, z) + LV(x, z), \quad f(z) \leqslant f(y) + \psi(z, y) + LV(z, y),$$

откуда

$$\psi(x, y) \leqslant f(x) - f(y) \leqslant \psi(x, z) + \psi(z, y) + LV(x, z) + LV(z, y),$$

т.е. верно (2.9) при $\psi_\delta = \psi$.

На (2.10) основан предложенный ранее проксимальный метод для вариационных неравенств [17]. Естественно возникает идея обобщить этот метод на абстрактные задачи (2.5) и (2.7) в предположениях их разрешимости, а также существования модели, удовлетворяющей определению 2.1.1. При этом будем учитывать погрешность δ в (2.9), а также погрешность $\widetilde{\delta}$ решения вспомогательных задач на итерациях согласно одному из достаточно известных в алгоритмической оптимизации подходов (см., например, раздел 3 из [16]):

$$x := \underset{y \in Q}{\overset{\widetilde{\delta}}{\arg\min}} \varphi(y), \text{ если } \langle \nabla\varphi(x), x - y\rangle \leqslant \widetilde{\delta} \quad \forall y \in Q. \tag{2.12}$$

Опишем $(N+1)$-ую итерацию рассматриваемого метода ($N = 0, 1, 2, \ldots$), выбрав начальное приближение $x^0 = \underset{x \in Q}{\arg\min}\, d(x)$, зафиксировав точность $\varepsilon > 0$, а также некоторую константу $L_0 \leqslant 2L$.

Замечание 2.1.5. Отметим, что в конце 2019 года на конференции NIPS-2019 (через несколько месяцев после опубликования в печати описываемых нами результатов) появилась работа [65], в которой предложен другой метод для вариационных неравенств (Algorithm 1 из [65]) с адаптивным выбором шага. Однако подход [65] к выбору шагов на итерациях отличен от нашего критерия выхода из итерации (2.13), а

Алгоритм 1 Адаптивный метод для вариационных неравенств, неточная модель.

1. $N := N + 1$; $L_{N+1} := \frac{L_N}{2}$.

2. Вычисляем:
$y^{N+1} := \underset{x \in Q}{\overset{\widetilde{\delta}}{\arg\min}} \left\{ \psi_\delta(x, x^N) + L_{N+1} V(x, x^N) \right\}$,
$x^{N+1} := \underset{x \in Q}{\overset{\widetilde{\delta}}{\arg\min}} \left\{ \psi_\delta(x, y^{N+1}) + L_{N+1} V(x, x^N) \right\}$
до тех пор, пока не будет выполнено:

$$\begin{aligned} \psi_\delta(x^{N+1}, x^N) \leqslant \psi_\delta(y^{N+1}, x^N) + \psi_\delta(x^{N+1}, y^{N+1}) + \\ + L_{N+1} V(y^{N+1}, x^N) + L_{N+1} V(x^{N+1}, y^{N+1}) + \delta. \end{aligned} \tag{2.13}$$

3. **Если** (2.13) не выполнено, **то** $L_{N+1} := 2L_{N+1}$ и повторяем п. 2.

4. **Иначе** переход к п. 1.

5. Критерий остановки метода:

$$S_N := \sum_{k=0}^{N-1} \frac{1}{L_{k+1}} \geqslant \frac{\max_{x \in Q} V(x, x^0)}{\varepsilon}. \tag{2.14}$$

также отсутствует какой-либо вариант адаптивного критерия остановки (2.14). Помимо этого, если говорить в терминах обозначений для алгоритма 1, методика выбора шага [65] связана с условием $L_k \leqslant L_{k+1}$ $(k = 0, 1, 2, ...)$, что по-видимому существенно ограничивает возможности адаптивной настройки при выборе шага. Действительно, потенциальная возможность малых значений L_k может позволить надеяться на скорое выполнение критерия остановки (2.14), что существенно обыгрывается далее при проведении вычислительных экспериментов (раздел 2.4) для негладких седловых задач. Во всяком случае, как показали результаты экспериментов для некоторых примеров, работа методов (достижение необходимого критерия остановки) очень сильно замедляется, если убрать деление на 2 в пункте 1 листинга алгоритма 1 (наглядно это описано далее в разделе 2.5, но замечалось и для других примеров).

При этом в [65] не рассматривались никакие вариации понятия неточного оракула для ВН (и, соответственно, этот подход не может приводить к универсальному методу; он ничего не позволяет утверждать для ВН с гёльдеровыми операторами), неточного проектирования (решения вспомогательных подзадач) на итерациях метода (2.12), композитных задач, а также условие относительной гладкости для ВН (см. пример 2.1.2). Однако представляется весьма интересным пример задачи о совместном использовании ресурсов из [65] с нелипшицевым оператором для ВН (пример 2.1.3 выше).

Справедлив следующий результат об оценке качества решения, выдаваемого алгоритмом 1.

Теорема 2.1.6. *Пусть ψ_δ есть (δ, L)-модель для задач* (2.5) *и* (2.7) *на множестве Q. Тогда после остановки рассматриваемого метода для всякого $x \in Q$ будет заведомо выполнено неравенство:*

$$-\frac{1}{S_N}\sum_{k=0}^{N-1}\frac{\psi_\delta(x, y^{k+1})}{L_{k+1}} \leqslant \frac{V(x, x^0)}{S_N} + 2\widetilde{\delta} + \delta \leqslant \varepsilon + 2\widetilde{\delta} + \delta, \tag{2.15}$$

а также

$$\psi(\widetilde{y}, x) \leqslant \frac{V(x, x^0)}{S_N} + 2\widetilde{\delta} + 3\delta \leqslant \varepsilon + 2\widetilde{\delta} + 3\delta \tag{2.16}$$

при

$$\widetilde{y} := \frac{1}{S_N}\sum_{k=0}^{N-1}\frac{y^{k+1}}{L_{k+1}}. \tag{2.17}$$

Доказательство. После завершения $(N+1)$-ой итерации метода ($N = 0, 1, 2\ldots$) ввиду (2.12) имеем:

$$\psi_\delta(y^{N+1}, x^N) \leqslant \psi_\delta(x^{N+1}, x^N) + L_{N+1}V(x^{N+1}, x^N) -$$
$$-L_{N+1}V(x^{N+1}, y^{N+1}) - L_{N+1}V(y^{N+1}, x^N) + \widetilde{\delta},$$
$$\psi(x^{N+1}, y^{N+1}) \leqslant \psi(x, y^{N+1}) + L_{N+1}V(x, x^N) -$$
$$-L_{N+1}V(x, x^{N+1}) - L_{N+1}V(x^{N+1}, x^N) + \widetilde{\delta}.$$

Далее, ввиду (2.13):

$$\psi_\delta(x^{N+1}, x^N) \leqslant \psi_\delta(y^{N+1}, x^N) + \psi_\delta(x^{N+1}, y^{N+1}) +$$
$$+L_{N+1}V(y^{N+1}, x^N) + L_{N+1}V(y^{N+1}, y^{N+1}) + \delta.$$

Отметим, что предположение (2.9) гарантирует выполнение условия (2.13) при $L_{N+1} \geqslant L$ после нескольких увеличений L_{N+1} в 2 раза. Просуммировав последние три неравенства, получаем:

$$\psi_\delta(x, y^{N+1}) + L_{N+1}V(x, x^N) - L_{N+1}V(x, x^{N+1}) \leqslant 2\widetilde{\delta} + \delta,$$

откуда

$$\frac{\psi_\delta(x, y^{N+1})}{L_{N+1}} + V(x, x^N) - V(x, x^{N+1}) \leqslant \frac{1}{L_{N+1}}(2\widetilde{\delta} + \delta),$$

или после суммирования:

$$-\sum_{k=0}^{N-1} \frac{\psi_\delta(x, y^{k+1})}{L_{k+1}} \leqslant \sum_{k=0}^{N-1} \left(V(x, x^k) - V(x, x^{k+1})\right) + (2\widetilde{\delta} + \delta)S_N,$$

откуда и следует доказываемое неравенство (2.15). □

Далее,

$$\psi(\widetilde{y}, x) \leqslant \frac{1}{S_N}\sum_{k=0}^{N-1} \frac{\psi(y^{k+1}, x)}{L_{k+1}} + \delta \leqslant -\frac{1}{S_N}\sum_{k=0}^{N-1} \frac{\psi_\delta(x, y^{k+1})}{L_{k+1}} + 2\delta \leqslant$$

$$\leqslant \frac{V(x, x^0)}{S_N} + 2\widetilde{\delta} + 3\delta,$$

что и требовалось.

Замечание 2.1.7. Ввиду (2.9) и выбора $L_0 \leqslant 2L$ гарантированно будет верно

$$L_{k+1} \leqslant 2L \ \ \forall k = \overline{0, N-1}.$$

Поэтому

$$S_N \geqslant \frac{N}{2L}$$

и (2.15)–(2.16) означают, что для всякого $x \in Q$ будут верны неравенства:

$$\psi_\delta(\widetilde{y}, x) \leqslant -\frac{1}{S_N}\sum_{k=0}^{N-1} \frac{\psi(x, y^{k+1})}{L_{k+1}} + 2\delta \leqslant \frac{2LV(x, x_0)}{N} + 2\widetilde{\delta} + 3\delta \leqslant \varepsilon + 2\widetilde{\delta} + 3\delta \tag{2.18}$$

после выполнения не более чем

$$O\left(\frac{1}{\varepsilon}\right) \tag{2.19}$$

итераций предлагаемого метода. При этом нетрудно проверить, количество решений вспомогательных задач в п. 2 алгоритма 1 на N итерациях метода не превышает

$$2N + \log_2 \frac{L}{L_0},$$

т.е. стоимость итерации в среднем будет сопоставимой со стоимостью итерации классического экстраградиентного метода, предполагающей решение двух вспомогательных задач на каждой итерации. Отметим, что оценка (2.19) с точностью до числового множителя оптимальна для вариационных неравенств и седловых задач [36, 103, 136, 182]. Предложенный алгоритм 1 применим и для более широкого класса задач равновесного программирования для функционала ψ, удовлетворяющего предположениям (i)–(iv) определения 2.1.1 выше и с точки зрения количества итераций будет оптимальным с точностью до числового множителя.

Замечание 2.1.8. Для обычных «слабых» вариационных неравенств (2.3) неравенство (2.18) можно заменить на

$$\max_{x \in Q} \langle G(x), \widetilde{y} - x \rangle \leqslant \varepsilon + 2\widetilde{\delta} + 3\delta. \tag{2.20}$$

Отметим, что именно (2.20) часто используют как критерий качества решения вариационного неравенства (см., например, [135]). При этом он может быть достигнут в случае монотонного G, даже если не существует точного решения x_*.

Замечание 2.1.9. Отметим, что если для всякого $x \in Q$ верно $\psi(x, x) = 0$, то (2.6) совпадает с известным условием кососимметричности для задач равновесного программирования, известного по работам А.С. Антипина, Ф.П. Васильева, а также их соавторов (см., например, [5–9]). Интересен вопрос о обобщении описанных нами результатов на задачи равновесного программирования с общими условиями кососимметричности ψ. Однако тут возникает проблема теоретически обосновать критерий качества решения вида (2.20) для точки $\widetilde{y}$ из (2.17), причём за соответствующее известным нижним оценкам количество итераций на рассматриваемых классах задач. Представляется, что это было бы интересно осмыслить и проработать в будущем.

2.2 Неточный оракул и универсальный метод для вариационных неравенств

Если в предыдущем пункте мы вели речь об общей концепции абстрактной неточной оптимизационной модели для вариационных неравенств и смежных задач, то сейчас остановимся на важных частных случаях. Прежде всего, покажем, как можно ввести для вариационных неравенств (уже без модельной общности) аналог (δ, L)-оракула для задач минимизации функционала (это позволяет применять разработанный метод к вариационным неравенствам с неточной информацией об операторе). Также будет показано, как можно предложить интерполяцию с искусственной неточностью для гёльдеровых операторов, что приводит к универсальному методу для вариационных неравенств с гёльдеровым монотонным оператором. Введем следующее определение неточного оракула для оператора G.

Определение 2.2.1. Пусть для некоторого $\delta_u > 0$ (неконтролируемая ошибка) и при любом $\delta_c > 0$ (контролируемая ошибка) существует константа $L(\delta_c) \in (0, +\infty)$ такая, что для всяких $x, y \in Q$ возможно вычислить $\widetilde{G}(x, \delta_c, \delta_u), \widetilde{G}(y, \delta_c, \delta_u) \in E^*$, для которых верны неравенства

$$\begin{aligned} &\langle \widetilde{G}(y, \delta_c, \delta_u) - \widetilde{G}(x, \delta_c, \delta_u), y - z \rangle \leqslant \\ &\leqslant \frac{L(\delta_c)}{2} \left(\|y - x\|^2 + \|y - z\|^2 \right) + \delta_c + \delta_u, \end{aligned} \tag{2.21}$$

$$\langle \widetilde{G}(y, \delta_c, \delta_u) - G(y), y - z \rangle \geqslant -\delta_u \quad \forall x, y, z \in Q. \tag{2.22}$$

Тогда оператор $\widetilde{G}(x, \delta_c, \delta_u)$ будем называть *неточным оракулом* ((δ, L)-оракулом) для оператора G.

В этом определении δ_c — ошибка оракула, которую возможно контролировать и выбирать, а δ_u — неконтролируемая ошибка. Рассмотрим некоторые примеры.

Пример 2.2.2 (*Липшицев оператор ВН с неточными значениями на ограниченном множестве*)**.** Предположим, что

1. Оператор G L-липшицев на Q, т.е. для всяких $x, y \in Q$ верно

$$\|G(x) - G(y)\|_* \leqslant L\|x - y\|.$$

2. Множество Q ограничено: $\max_{x,y\in Q}\|x-y\|\leqslant D$.

3. Существует $\bar{\delta}_u>0$ и в любой точке $x\in Q$ возможно вычислить приближение $\widetilde{G}(x)$ для $G(x)$ с точностью $\|\widetilde{G}(x)-G(x)\|_*\leqslant\bar{\delta}_u$.

Тогда для всякого $z\in Q$ верно

$$\langle\widetilde{G}(y)-\widetilde{G}(x),y-z\rangle=$$

$$=\langle\widetilde{G}(y)-G(y),y-z\rangle-\langle\widetilde{G}(x)-G(x),y-z\rangle+\langle G(y)-G(x),y-z\rangle\leqslant$$

$$\leqslant 2\bar{\delta}_uD+\|G(y)-G(x)\|_*\|y-z\|\leqslant 2\bar{\delta}_uD+L\|y-x\|\|y-z\|\leqslant$$

$$\leqslant 2\bar{\delta}_uD+\frac{L}{2}\left(\|y-x\|^2+\|y-z\|^2\right).$$

Таким образом, можно положить в определении 2.2.1 $\delta_u=2\bar{\delta}_uD$ и $L(\delta_c)\equiv L$. Тогда

$$|\langle\widetilde{G}(y)-G(y),y-z\rangle|\leqslant\|\widetilde{G}(y)-G(y)\|_*\|y-z\|\leqslant\bar{\delta}_uD=\frac{1}{2}\delta_u<\delta_u,$$

откуда

$$\langle\widetilde{G}(y)-G(y),y-z\rangle>-\delta_u.$$

Итак, при выбранных значениях параметров $\widetilde{G}(x)=\widetilde{G}(x,\delta_c,\delta_u)$ удовлетворяет определению 2.2.1.

Отметим важный вспомогательный результат, который позволяет ввести оптимизационную модель (интерполяцию) для вариационных неравенств с гёльдеровыми монотонными операторами, на базе которой можно ввести в рассмотрение универсальный метод для ВН.

Лемма 2.2.3. *Пусть $a,b,c\geqslant 0$, $\nu\in[0,1]$. Тогда для любого $\delta>0$,*

$$ab^\nu c\leqslant\left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}}\frac{a^{\frac{2}{1+\nu}}}{2}\left(b^2+c^2\right)+\frac{\delta}{2}.$$

Доказательство. Зафиксируем некоторое $\nu\in[0,1]$. Тогда для всякого $x\in[0,1]$ верно $x^{2\nu}\leqslant 1$. С другой стороны, при $x\geqslant 1$ верно $x^{2\nu}\leqslant x^2$. Поэтому для произвольного $x\geq 0$ имеем $x^{2\nu}\leqslant x^2+1$. Это означает, что при $\alpha,\beta\geqslant 0$

$$\alpha^\nu\beta\leqslant\frac{\alpha^{2\nu}}{2}+\frac{\beta^2}{2}\leqslant\frac{\alpha^2}{2}+\frac{\beta^2}{2}+\frac{1}{2}.$$

Выбрав $\alpha = \frac{ba^{\frac{1}{1+\nu}}}{\delta^{\frac{1}{1+\nu}}}$ и $\beta = \frac{ca^{\frac{1}{1+\nu}}}{\delta^{\frac{1}{1+\nu}}}$, имеем:

$$\frac{b^\nu a^{\frac{\nu}{1+\nu}}}{\delta^{\frac{\nu}{1+\nu}}} \frac{ca^{\frac{1}{1+\nu}}}{\delta^{\frac{1}{1+\nu}}} \leqslant \frac{b^2 a^{\frac{2}{1+\nu}}}{2\delta^{\frac{2}{1+\nu}}} + \frac{c^2 a^{\frac{2}{1+\nu}}}{2\delta^{\frac{2}{1+\nu}}} + \frac{1}{2},$$

а также

$$ab^\nu c \leqslant \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} \frac{a^{\frac{2}{1+\nu}}}{2} \left(b^2 + c^2\right) + \frac{\delta}{2}.$$

□

Пример 2.2.4 (*Применимость понятия неточного оракула для ВН в случае гёльдерова оператора*)**.** Предположим, что оператор G удовлетворяет условию Гёльдера на Q, то есть

$$\|G(x) - G(y)\| \leqslant L_\nu \cdot \|x - y\|^\nu \quad \forall x, y \in Q \tag{2.23}$$

и некоторого $\nu \in [0; 1]$.

Применяя лемму 2.2.3, получаем для любых $x, y, z \in Q$ и $\delta > 0$:

$$\langle G(y) - G(x), y - z\rangle \leqslant \|G(y) - G(x)\|_* \|y - z\| \leqslant L_\nu \|x - y\|^\nu \|y - z\| \leqslant$$

$$\leqslant \frac{1}{2}\left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}} \left(\|x - y\|^2 + \|y - z\|^2\right) + \frac{\delta}{2}.$$

Таким образом, если выбрать $\delta_u = 0$, $\delta_c = \frac{\delta}{2}$, а также $L(\delta_c) = \left(\frac{1}{2\delta_c}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$, то будет верно (2.21).

Таким образом, в случае ν-гёльдерова оператора поля G (удовлетворяющего (2.2)) верно неравенство

$$\langle G(z) - G(y), z - x\rangle \leqslant \frac{L}{2}||z - x||^2 + \frac{L}{2}||z - y||^2 + \frac{\varepsilon}{2} \tag{2.24}$$

для некоторой константы $L = \left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} L^{\frac{2}{1+\nu}}$. На базе интерполяции (2.24) алгоритм 1 сводится к следующему *универсальному* методу для вариационных неравенств [17].

Предположим, что оператор G монотонен и ν-гёльдеров для некоторого $\nu \in [0; 1]$. Тогда из теоремы 2.1.6 и неравенства (2.24) следует, что после N итераций алгоритма 2 для всякого $x \in Q$ верно неравенство

$$\langle G(x), \widetilde{y} - x\rangle \leqslant \frac{L^{\frac{2}{1+\nu}}}{\varepsilon^{\frac{1-\nu}{1+\nu}}} \frac{V(x, x^0)}{N} + \frac{3\varepsilon}{2},$$

Алгоритм 2 Универсальный метод для вариационных неравенств.

1. $N := N + 1$; $L_{N+1} := \frac{L_N}{2}$.

2. Вычисляем:
$$y^{N+1} := \arg\min_{x \in Q}^{\widetilde{\delta}} \left\{ \left\langle G(x^N), x - x^N \right\rangle + L_{N+1} V(x, x^N) \right\},$$
$$x^{N+1} := \arg\min_{x \in Q}^{\widetilde{\delta}} \left\{ \left\langle G(y^{N+1}), x - y^{N+1} \right\rangle + L_{N+1} V(x, x^N) \right\}$$
до тех пор, пока не будет выполнено:
$$\left\langle G(y^{N+1}) - G(x^N), y^{N+1} - x^{N+1} \right\rangle \leqslant$$
$$\leqslant L_{N+1} V(y^{N+1}, x^N) + L_{N+1} V(x^{N+1}, y^{N+1}) + \delta.$$

3. **Если** (2.13) не выполнено, **то** $L_{N+1} := 2L_{N+1}$ и повторяем п. 2.

4. **Иначе** переход к п. 1.

5. Критерий остановки метода:
$$S_N := \sum_{k=0}^{N-1} \frac{1}{L_{k+1}} \geqslant \frac{\max_{x \in Q} V(x, x^0)}{\varepsilon}.$$

откуда при условии

$$N \geqslant \frac{2L^{\frac{2}{1+\nu}}}{\varepsilon^{\frac{2}{1+\nu}}} \max_{x \in Q} V(x, x^0) \tag{2.25}$$

гарантированно будет верно $\max\limits_{x \in Q} \langle G(x), \widetilde{y} - x \rangle \leqslant 2\varepsilon$.

Однако за счёт адаптивного критерия остановки предложенный метод (частный случай алгоритма 1) может потенциально позволить достичь приемлемого качества решения за меньшее по сравнению с (2.25) число итераций. Более того, предполагает адаптивную настройку на оптимальный параметр $\nu \in [0; 1]$ (уровень гладкости) оператора G. Если считать, что условие Гёльдера (2.2) верно для произвольного $\nu \in [0; 1]$ и $L_0 < +\infty$ (другие константы L_ν $(\nu \neq 0)$ могут быть бесконечными),

то доказанная оценка (2.25) примет вид

$$\inf_{\nu\in[0;1]} \frac{2L^{\frac{2}{1+\nu}}}{\varepsilon^{\frac{2}{1+\nu}}} \max_{x\in Q} V(x, x^0),$$

а оптимальный параметр ν для соответствующей задачи по сути подбирается на итерациях алгоритма 2 при адаптивной настройке локальных констант L_{k+1} $(k = 0, 1, ...)$. Оказывается, что настройка метода на уровень гладкости оператора ВН может позволить улучшить качество работы метода по сравнению с теоретическими вычислительными гарантиями (они заведомо оптимальны при $\nu = 0$ и $\nu = 1$). Например, для задачи с ограниченным ВН $(\nu = 0)$ метод на практике может приводить к скорости работы $O\left(\frac{1}{\varepsilon}\right)$, оптимальной для задач с липшицевым оператором $(\nu = 1)$ и даже выше. Далее, в разделе 2.4 будут приведены вычислительные эксперименты для некоторых примеров таких задач.

Заметим также, что полученная выше теорема 2.1.6 позволяет обобщить этот подход на смешанные вариационные неравенства (пример 2.0.2) с гёльдеровым оператором G.

Отметим еще пример, связанный с комбинацией естественной и искусственной неточностей задания вариационных неравенств.

Пример 2.2.5 (*Гёльдеров оператор с неточным заданием оператора на ограниченном допустимом множестве*)**.** Пусть

1. Оператор $G(x)$ гёльдеров на Q, т.е. для некоторых $\nu \in [0, 1]$ и $L_\nu \geqslant 0$
$$\|G(x) - G(y)\|_* \leqslant L_\nu \|x - y\|^\nu, \quad x, y \in Q.$$
2. Множество Q ограничено: $\max_{x,y\in Q} \|x - y\| \leqslant D$.
3. Существует $\bar{\delta}_u > 0$ и в любой точке $x \in Q$, возможно вычислить приближение $\widetilde{G}(x)$ для $G(x)$ то есть $\|\widetilde{G}(x) - G(x)\|_* \leqslant \bar{\delta}_u$.

Тогда для всякого $z \in Q$ верны неравенства

$$\begin{aligned}\langle \widetilde{G}(y) - \widetilde{G}(x), y - z\rangle &= \langle \widetilde{G}(y) - G(y), y - z\rangle - \\ -\langle \widetilde{G}(x) - G(x), y - z\rangle &+ \langle G(y) - G(x), y - z\rangle \leqslant \\ &\leqslant 2\bar{\delta}_u D + \|G(y) - G(x)\|_* \|y - z\| \leqslant \\ &\leqslant 2\bar{\delta}_u D + L_\nu \|y - x\|^\nu \|y - z\| \leqslant\end{aligned}$$

$$\leqslant 2\bar{\delta}_u D + \frac{1}{2}\left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}} \left(\|x-y\|^2 + \|y-z\|^2\right) + \frac{\delta}{2}$$

согласно лемме 2.2.3.

Таким образом, если положить $\delta_u = 2\bar{\delta}_u D$, а также $\delta_c = \frac{\delta}{2}$ и

$$L(\delta_c) = \left(\frac{1}{2\delta_c}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}},$$

то будет верно (2.21). Неравенство (2.22) можно проверить аналогично примеру 2.2.2. Итак, при подобранных значениях параметров $\widetilde{G}(x) = \widetilde{G}(x, \delta_c, \delta_u)$ удовлетворяет определению 2.2.1.

Теперь проиллюстрируем связь введенного нами понятия неточной модели (см. определение 2.2.1) для вариационных неравенств с известным понятием (δ, L)-оракула для задач минимизации функционалов.

Пример 2.2.6 (*Связь с концепцией (δ, L)-оракула для задач минимизации функционалов*)**.** Пусть выпуклая функция $f : Q \to \mathbb{R}$, где Q — выпуклый компакт. Пусть f допускает (δ, L)-оракул в любой точке, то есть для некоторого $L > 0$ в любом $y \in Q$ существует пара $(f_\delta(y), g_\delta(y)) \in \mathbb{R} \times \mathbb{R}^n$, такая, что для всех $y \in Q$

$$f_\delta(y) + \langle g_\delta(y), x - y\rangle \leqslant f(x) \leqslant f_\delta(y) + \langle g_\delta(y), x - y\rangle + \frac{L\|x-y\|^2}{2} + \delta \quad (2.26)$$

Покажем, что g_δ удовлетворяет определению 2.2.1. Ясно, что для произвольного $x \in Q$ $f_\delta(x) \leqslant f(x) \leqslant f_\delta(x) + \delta$. Далее, согласно левому неравенству в (2.26):

$$f(x) \geqslant f_\delta(y) + \langle g_\delta(y), x - y\rangle \geqslant f(y) + \langle g_\delta(y), x - y\rangle - \delta,$$

и

$$\langle g_\delta(y), x - y\rangle \leqslant f(x) - f(y) + \delta \quad \forall\, x, y \in Q. \quad (2.27)$$

Неравенство справа в (2.26) означает, что для всяких $x, y, z \in Q$:

$$f(x) \leqslant f_\delta(z) + \langle g_\delta(z), x - z\rangle + \frac{L}{2}\|x-z\|^2 + \delta \leqslant$$

$$\leqslant f(z) + \langle g_\delta(z), x - z\rangle + \frac{L}{2}\|x-z\|^2 + \delta,$$

и

$$f(x) - f(z) \leqslant \langle g_\delta(z), x - z\rangle + \frac{L}{2}\|x-z\|^2 + \delta.$$

Аналогично

$$f(z) - f(y) \leqslant \langle g_\delta(y), z - y\rangle + \frac{L}{2}\|z - y\|^2 + \delta.$$

Поэтому $f(x) - f(y) - \langle g_\delta(z), x - z\rangle - \langle g_\delta(y), z - y\rangle \leqslant \frac{L}{2}\|x - z\|^2 + \frac{L}{2}\|z - y\|^2 + 2\delta$. Кроме того, (2.27) означает:

$$\langle g_\delta(y), x-y\rangle - \langle g_\delta(z), x-z\rangle - \langle g_\delta(y), z-y\rangle \leqslant \frac{L}{2}\|z-x\|^2 + \frac{L}{2}\|y-z\|^2 + 3\delta, \text{ т.е.}$$

$$\langle g_\delta(y) - g_\delta(z), x - z\rangle \leqslant \frac{L}{2}\|z - x\|^2 + \frac{L}{2}\|y - z\|^2 + 3\delta,$$

или

$$\langle g_\delta(z) - g_\delta(y), z - x\rangle \leqslant \frac{L}{2}\|z - x\|^2 + \frac{L}{2}\|z - y\|^2 + 3\delta, \quad \forall\, x, y, z \in Q.$$

Таким образом, (2.21) верно при $L(\delta_c) = L$, $\delta_c = 0$ и $\delta_u = 3\delta$, а также $\widetilde{g}(y, \delta_c, \delta_u) = g_\delta(y)$. При дополнительном предположении $\|g_\delta(y) - G(y)\| \leqslant \bar{\delta}_u$ получаем:

$$\langle g_\delta(y) - G(y), y - x\rangle \geqslant -\bar{\delta}_u\|x - y\| \geqslant -\bar{\delta}_u D, \quad \forall x, y \in Q,$$

т.е. верно (2.22).

2.3 Понятие (δ, L)-модели функции для седловых задач и оценки скорости сходимости предложенного алгоритма

Вариационные неравенства с монотонными операторами возникают, в частности, при решении выпукло-вогнутых седловых задач. В таких задачах для выпуклого по u и вогнутого по v функционала $f(u, v) : \mathbb{R}^{n_1+n_2} \to \mathbb{R}$ ($u \in Q_1 \subset \mathbb{R}^{n_1}$ и $v \in Q_2 \subset \mathbb{R}^{n_2}$) требуется найти (u_*, v_*), такую, что:

$$f(u_*, v) \leqslant f(u_*, v_*) \leqslant f(u, v_*) \tag{2.28}$$

для произвольных $u \in Q_1$ и $v \in Q_2$. Мы считаем Q_1 и Q_2 выпуклыми компактами в пространствах $\mathbb{R}^{n_1}$ и $\mathbb{R}^{n_2}$ и поэтому $Q = Q_1 \times Q_2 \subset \mathbb{R}^{n_1+n_2}$ также есть выпуклый компакт. Для всякого $x = (u, v) \in Q$ будем полагать, что

$$||x|| = \sqrt{||u||_1^2 + ||v||_2^2},$$

где $||\cdot||_1$ и $||\cdot||_2$ — нормы в пространствах $\mathbb{R}^{n_1}$ и $\mathbb{R}^{n_2}$). Будем обозначать $x = (u_x, v_x),\ y = (u_y, v_y) \in Q$.

Хорошо известно, что для достаточно гладкой функции f по u и v задача (2.28) сводится к вариационному неравенству с оператором

$$G(x) = \begin{pmatrix} f'_u(u_x, v_x) \\ -f'_v(u_x, v_x) \end{pmatrix}. \tag{2.29}$$

В силу выпуклости f в u и вогнутости в v, оператор G является монотонным $\langle G(x) - G(y), x - y\rangle \geqslant 0 \quad \forall x, y \in Q \subset E$, где $x = (u_1, v_1), y = (u_2, v_2)$.

Предложим некоторый вариант концепции (δ, L)-модели, применимую для седловых задач.

Определение 2.3.1. Будем говорить, что функция $\psi_\delta(x, y)$ ($\psi : \mathbb{R}^{n_1+n_2} \times \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}$) есть (δ, L)-**модель** для седловой задачи (2.28) на множестве Q, если для функционала ψ при всяких $x, y, z \in Q$ выполнены предположения:

(i) функционал $\psi_\delta(x, y)$ выпуклый по первой переменной;

(ii) $\psi_\delta(x, x) = 0 \ \ \forall x \in Q$;

(iii) *абстрактная δ-монотонность*;

(iv) (*обобщенная гладкость*)

$$\psi_\delta(x, y) \leqslant \psi_\delta(x, z) + \psi_\delta(z, y) + LV(x, z) + LV(z, y) + \delta$$

для некоторой фиксированной постоянной $L > 0$, где $\delta > 0$ — некоторая постоянная величина (оценка погрешности задания ψ, степень отклонения от гладкости);

(v) справедливо неравенство:

$$f(u_y, v_x) - f(u_x, v_y) \leqslant -\psi_\delta(x, y) \quad \forall x, y \in Q.$$

Пример 2.3.2. Предложенная концепция модели функции для седловых задач вполне применима, например для рассмотренных в статье [84] композитных седловых задач вида:

$$f(u, v) = \widetilde{f}(u, v) + h(u) - \varphi(v)$$

для некоторой выпуклой по u и вогнутой по v субдифференцируемой функции $\widetilde{f}$, а также выпуклых функций простой структуры h и φ (для этих функций операция проектирования на множество не очень затратна). В таком случае можно положить

$$\psi(x,y) = \langle G(y), x-y\rangle + h(u_x) + \varphi(v_x) - h(u_y) - \varphi(v_y),$$

где

$$G(y) = \begin{pmatrix} \widetilde{f}'_u(u_y,v_y) \\ -\widetilde{f}'_v(u_y,v_y) \end{pmatrix}.$$

Действительно, из субградиентных неравенств получаем:

$$\widetilde{f}(u_y,v_y) - \widetilde{f}(u_x,v_y) \leqslant \langle -\widetilde{f}'_u(u_y,v_y), u_x - u_y\rangle,$$

$$\widetilde{f}(u_y,v_x) - \widetilde{f}(u_y,v_y) \leqslant \langle -\widetilde{f}'_v(u_y,v_y), v_x - v_y\rangle.$$

Поэтому имеем

$$\widetilde{f}(u_y,v_x) - \widetilde{f}(u_x,v_y) \leqslant -\langle G(y), x-y\rangle,$$

откуда

$$f(u_y,v_x) - f(u_x,v_y) = \widetilde{f}(u_y,v_x) + h(u_y) - \varphi(v_x) - \widetilde{f}(u_x,v_y) - h(v_x) + \varphi(v_y) =$$

$$= \widetilde{f}(u_y,v_x) - \widetilde{f}(u_x,v_y) + h(u_y) + \varphi(v_y) - h(v_x) - \varphi(v_x) \leqslant$$

$$\leqslant -\langle G(y), x-y\rangle + h(u_y) + \varphi(v_y) - h(v_x) - \varphi(v_x) = -\psi_\delta(x,y).$$

Из теоремы 2.1.6 вытекает

Теорема 2.3.3. *Если для седловой задачи* (2.28) *существует* (δ, L)*-модель* $\psi(x,y)$*, то после остановки алгоритма 2 получаем точку*

$$\widetilde{y} = (u_{\widetilde{y}}, v_{\widetilde{y}}) := (\widetilde{u}, \widetilde{v}) := \frac{1}{S_N}\sum_{k=0}^{N_1} \frac{y_{k+1}}{L_{k+1}}, \tag{2.30}$$

для которой верна оценка величины-качества решения седловой задачи:

$$\max_{v\in Q_2} f(\widetilde{u}, v) - \min_{u\in Q_1} f(u, \widetilde{v}) \leqslant \varepsilon + 2\widetilde{\delta} + 2\delta. \tag{2.31}$$

Ясно, что универсальный метод для ВН можно применить и к выпукло-вогнутым седловым задачам. По-прежнему рассматриваем задачу нахождения седловой точки

$$f^* = \min_{u \in Q_1} \max_{v \in Q_2} f(u, v), \tag{2.32}$$

где $Q_1 \subset E_1$ и $Q_2 \subset E_2$ являются выпуклыми и замкнутыми подмножествами нормированных пространств E_1 и E_2 с нормами $\|\cdot\|_1$ и $\|\cdot\|_2$ соответственно. Основываясь на нормах в E_1 и E_2, мы определяем норму их произведения $E_1 \times E_2$ как $\|x\| = \max\{\|u\|_1, \|v\|_2\}$, $x = (u, v) \in E_1 \times E_2$ с соответствующей двойной нормой $\|s\|_* = \|z\|_{1,*} + \|w\|_{2,*}$, $s = (z, w) \in E^*$, где $\|\cdot\|_{1,*}$ и $\|\cdot\|_{2,*}$ — нормы на сопряженных пространствах E_1^* и E_2^*, двойственных к $\|\cdot\|_1$ и $\|\cdot\|_2$ соответственно.

Следующая лемма описывает достаточные условия, гарантирующие выполнение условия Гёльдера для оператора G из (2.29).

Лемма 2.3.4. *Предположим, что f в (2.32) таков, что существует число $\nu \in [0, 1]$ и константы $L_{11,\nu}, L_{12,\nu}, L_{21,\nu}, L_{22,\nu} < +\infty$, для которых*

$$\|\nabla_u f(u + \Delta u, v + \Delta v) - \nabla_u f(u, v)\|_{1,*} \leqslant L_{11,\nu}\|\Delta u\|_1^\nu + L_{12,\nu}\|\Delta v\|_2^\nu, \tag{2.33}$$

$$\|\nabla_v f(u + \Delta u, v + \Delta v) - \nabla_v f(u, v)\|_{2,*} \leqslant L_{21,\nu}\|\Delta u\|_1^\nu + L_{22,\nu}\|\Delta v\|_2^\nu \tag{2.34}$$

для всех $u, u + \Delta u \in Q_1, v, v + \Delta v \in Q_2$. Тогда G из (2.29) гёльдеров, то есть удовлетворяет (2.23) с теми же ν и $L_\nu = L_{11,\nu} + L_{12,\nu} + L_{21,\nu} + L_{22,\nu}$.

Доказательство. Действительно, для каждого $x = (u_1, v_1), y = (u_2, v_2) \in Q$ имеем:

$$\|G(x) - G(y)\|_* =$$

$$= \|\nabla_u f(u_1, v_1) - \nabla_u f(u_2, v_2)\|_{1,*} + \|\nabla_v f(u_1, v_1) - \nabla_v f(u_2, v_2)\|_{2,*} \leqslant$$

$$\leqslant L_{11,\nu}\|u_1 - u_2\|_1^\nu + L_{12,\nu}\|v_1 - v_2\|_2^\nu + L_{21,\nu}\|u_1 - u_2\|_1^\nu + L_{22,\nu}\|v_1 - v_2\|_2^\nu =$$

$$= (L_{11,\nu} + L_{21,\nu})\|u_1 - u_2\|_1^\nu + (L_{12,\nu} + L_{22,\nu})\|v_1 - v_2\|_2^\nu \leqslant$$

$$\leqslant (L_{11,\nu} + L_{12,\nu} + L_{21,\nu} + L_{22,\nu}) \max\{\|u_1 - u_2\|_1^\nu, \|v_1 - v_2\|_2^\nu\} =$$

$$= (L_{11,\nu} + L_{12,\nu} + L_{21,\nu} + L_{22,\nu})\|x - y\|^\nu.$$

□

Замечание 2.3.5. В качестве альтернативного подхода можно рассмотреть следующие основные и двойственные нормы в пространствах $E = E_1 \times E_2$ $\|x\| = \sqrt{\|u\|_1^2 + \|v\|_2^2}$, $x = (u, v) \in E_1 \times E_2$, и $\|s\|_* = \sqrt{\|z\|_{1,*}^2 + \|w\|_{2,*}^2}$, $s = (z, w) \in E^*$, где $\|\cdot\|_{1,*}$ и $\|\cdot\|_{2,*}$ нормы на $(E_1^*, \|\cdot\|_1)$ и $E_2^*, \|\cdot\|_2$ пространствах E_1^* и E_2^*, двойственных к $\|\cdot\|_1$ и $\|\cdot\|_2$ соответственно. В таком случае для всяких $x = (u_1, v_1), y = (u_2, v_2) \in Q$ имеем:

$$\|G(x) - G(y)\|_*^2 =$$

$$= \|\nabla_u f(u_1, v_1) - \nabla_u f(u_2, v_2)\|_{1,*}^2 + \|\nabla_v f(u_1, v_1) - \nabla_v f(u_2, v_2)\|_{2,*}^2 \leqslant$$

$$\leqslant 2(L_{11,\nu}^2 \|u_1 - u_2\|_1^{2\nu} + L_{12,\nu}^2 \|v_1 - v_2\|_2^{2\nu} + L_{21,\nu}^2 \|u_1 - u_2\|_1^{2\nu} +$$

$$+ L_{22,\nu}^2 \|v_1 - v_2\|_2^{2\nu}) =$$

$$= 2(L_{11,\nu}^2 + L_{21,\nu}^2)\|u_1 - u_2\|_1^{2\nu} + (L_{12,\nu}^2 + L_{22,\nu}^2)\|v_1 - v_2\|_2^{2\nu} \leqslant$$

$$\leqslant (L_{11,\nu}^2 + L_{12,\nu}^2 + L_{21,\nu}^2 + L_{22,\nu}^2) \max\{\|u_1 - u_2\|_1^{2\nu}, \|v_1 - v_2\|_2^{2\nu}\}) \leqslant$$

$$\leqslant 2(L_{11,\nu}^2 + L_{12,\nu}^2 + L_{21,\nu}^2 + L_{22,\nu}^2)\|x - y\|^{2\nu}$$

и

$$\|G(x) - G(y)\|_* \leqslant \sqrt{2(L_{11,\nu}^2 + L_{12,\nu}^2 + L_{21,\nu}^2 + L_{22,\nu}^2)}\|x - y\|^{\nu}.$$

Замечание 2.3.6. Вообще говоря, если множество Q ограничено, можно рассмотреть различный уровень гладкости в (2.33) и (2.34). Предположим, что для некоторых чисел $\nu_{11}, \nu_{12}, \nu_{21}, \nu_{22} \in [0; 1]$:

$$\|\nabla_u f(u + \Delta u, v + \Delta v) - \nabla_u f(u, v)\|_{1,*} \leqslant \widehat{L}_{11}\|\Delta u\|_1^{\nu_{11}} + \widehat{L}_{12}\|\Delta v\|_2^{\nu_{12}}, \quad (2.35)$$

$$|\nabla_v f(u + \Delta u, v + \Delta v) - \nabla_v f(u, v)\|_{2,*} \leqslant \widehat{L}_{21}\|\Delta u\|_1^{\nu_{21}} + \widehat{L}_{22}\|\Delta v\|_2^{\nu_{22}} \quad (2.36)$$

для всех $u, u + \Delta u \in Q_1, v, v + \Delta v \in Q_2$. Тогда утверждение леммы 2.3.4 заведомо верно выполняется при $\nu = \min\{\nu_{11}, \nu_{12}, \nu_{21}\, \nu_{22}\} \in [0; 1]$. Действительно, (2.35) и (2.36) означают, что

$$\|\nabla_u f(u + \Delta u, v + \Delta v) - \nabla_u f(u, v)\|_{1,*} \leqslant$$

$$\leqslant \widehat{L}_{11} \cdot D_Q^{\nu_{11} - \nu} \cdot \|\Delta u\|_1^{\nu} + \widehat{L}_{12} \cdot D_Q^{\nu_{12} - \nu} \cdot \|\Delta v\|_2^{\nu},$$

$$\|\nabla_v f(u + \Delta u, v + \Delta v) - \nabla_v f(u, v)\|_{2,*} \leqslant$$

$$\leqslant \widehat{L}_{21} \cdot D_Q^{\nu_{21} - \nu} \cdot \|\Delta u\|_1^{\nu} + \widehat{L}_{22} \cdot D_Q^{\nu_{22} - \nu} \cdot \|\Delta v\|_2^{\nu}$$

для всех $u, u + \Delta u \in Q_1, v, v + \Delta v \in Q_2$, $D_Q = \sup\{\|x - y\| \,|\, x, y \in Q\}$.

Следующая теорема показывает применимость алгоритма 2 для решения седловых задач вида (2.32).

Теорема 2.3.7. *Пусть выполнены предположения леммы 2.3.4 и множество Q ограничено. Предположим также, что выполнены k итераций алгоритма 2 с точностью ε для оператора G из (2.29) и $y^i = (u^i, v^i)$ — последовательность, сгенерированная этим алгоритмом. В таком случае верно неравенство:*

$$\max_{v\in Q_2} f(\widetilde{u}, v) - \min_{u\in Q_1} f(u, \widetilde{v}) \leqslant \frac{2L_\nu^{\frac{2}{1+\nu}}}{k\varepsilon^{\frac{1-\nu}{1+\nu}}} \max_{x\in Q} V(x, x^0) + \frac{\varepsilon}{2},$$

$$\text{где} \quad (\widetilde{u}, \widetilde{v}) = \frac{1}{S_k}\sum_{i=0}^{k-1} \frac{(u^i, v^i)}{L_{i+1}}, \quad S_k = \sum_{i=0}^{k-1} \frac{1}{L_{i+1}}.$$

Таким образом, за

$$O\left(\inf_{\nu\in[0,1]} \left(\frac{L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}} \cdot \max_{x\in Q} V(x, x^0)\right)$$

итераций алгоритм 2 гарантированно выдаст пару $y = (\widetilde{u}, \widetilde{v})$, для которой верно неравенство

$$\max_{v\in Q_2} f(\widetilde{u}, v) - \min_{u\in Q_1} f(u, \widetilde{v}) \leqslant \varepsilon.$$

Доказательство. Ввиду выпуклости f по переменной u и вогнутости f по переменной v для всякого $u \in Q_1$ имеем

$$\frac{1}{S_k}\sum_{i=0}^{k-1} \left\langle \nabla_u f(u^i, v^i), u^i - u\right\rangle_1 \geqslant \frac{1}{S_k}\sum_{i=0}^{k-1} \frac{1}{L_{i+1}}(f(u^i, v^i) - f(u, v^i)) \geqslant$$

$$\geqslant \frac{1}{S_k}\sum_{i=0}^{k-1} \frac{1}{L_{i+1}} f(u^i, v^i) - f(u, \widetilde{v}).$$

Аналогично для всякого $v \in Q_2$ получаем

$$\frac{1}{S_k}\sum_{i=0}^{k-1} \frac{1}{L_{i+1}} \left\langle -\nabla_v f(u^i, v^i), v^i - v\right\rangle_2 \geqslant -\frac{1}{S_k}\sum_{i=0}^{k-1} \frac{1}{L_{i+1}} f(u^i, v^i) + f(\widetilde{u}, v).$$

Просуммировав эти неравенства с учетом (2.30) и (2.31), получаем, что для произвольных $u \in Q_1$, $v \in Q_2$

$$f(\widetilde{u}, v) - f(u, \widetilde{v}) \leqslant \frac{1}{S_k} \sum_{i=0}^{k-1} \frac{1}{L_{i+1}} \langle G(y^i), y^i - x \rangle \leqslant \frac{1}{\sum_{i=0}^{k-1} \frac{1}{L_{i+1}}} V(x, x^0) + \frac{\varepsilon}{2}.$$

Далее, $L_{i+1} \leqslant 2L\left(\frac{\varepsilon}{2}\right)$, где $L\left(\frac{\varepsilon}{2}\right) = \left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} L_{\frac{2}{1+\nu}}$ и множество Q ограничено. Поэтому

$$\max_{v \in Q_2} f(\widetilde{u}, v) - \min_{u \in Q_1} f(u, \widetilde{v}) \leqslant \frac{2L_\nu^{\frac{2}{1+\nu}}}{k\varepsilon^{\frac{1-\nu}{1+\nu}}} \max_{x \in Q} V(x, x^0) + \frac{\varepsilon}{2},$$

откуда и следует доказываемое утверждение. □

Замечание 2.3.8. Для седловой точки $(u_*, v_*) \in Q$ верно $\max\limits_{v \in Q_2} f(u_*, v) = \min\limits_{u \in Q_1} f(u, v_*)$. Поэтому неравенство $\max_{v \in Q_2} f(\widetilde{u}, v) - \min_{u \in Q_1} f(u, \widetilde{v}) \leqslant \varepsilon$, означает, что $(\widetilde{u}, \widetilde{v})$ — достаточно приемлемое приближение для (u_*, v_*).

Важным частным случаем седловой задачи является задача нахождения решения лагранжевых седловых задач для выпуклого программирования. Рассмотрим следующую задачу выпуклого программирования:

$$\widehat{f}(x) \to \min,\ x \in Q, \quad \varphi_j(x) \leqslant 0, \quad j = 1, ..., m, \tag{2.37}$$

где Q — компакт, f и φ_j — выпуклые функционалы с гёльдеровыми градиентами (или с ограниченными субградиентами в случае $\nu_j = 0$)

$$\|\nabla \widehat{f}(x) - \nabla \widehat{f}(y)\|_* \leqslant L_{\nu_0} \|x - y\|^{\nu_0},$$

$$\|\nabla \varphi_j(x) - \nabla \varphi_j(y)\|_* \leqslant L_{\nu_j} \|x - y\|^{\nu_j} \quad \forall x, y \in Q, j = 1, ..., m$$

для некоторых $\nu_0, ..., \nu_m \geqslant 0$ и $L_{\nu_0}, ..., L_{\nu_m} > 0$. Соответствующая функция Лагранжа для задачи (2.37) имеет вид $L(x, \lambda) = \widehat{f}(x) + \sum\limits_{j=1}^{m} \lambda_j \varphi_j(x)$, где $\lambda_j \geqslant 0$ $(j = 1, ..., m)$ множители Лагранжа. Если точка (x_*, λ_*) — седловая точка выпукло-вогнутой функции Лагранжа $L(x, \lambda)$, то x_* — решение (2.37). Предположим также, что для задачи выпуклого программирования (2.37) выполнено условие Слейтера, т.е. существует точка $\bar{x}$, для которой $\varphi_j(\bar{x}) < 0$, $j = 1, ..., m$. Тогда можно показать, что вектор

оптимальных множителей Лагранжа λ_* лежит в некотором ограниченном допустимом множестве. Таким образом, вместо задачи минимизации (2.37) можно рассмотреть седловую задачу $\min_{x\in Q}\max_{\lambda\in\Lambda} L(x,\lambda)$, которая является выпукло-вогнутой задачей на ограниченном множестве. Использование леммы 2.3.4 и предположения о гёльдеровости (суб)градиентов $\widehat{f}$ и φ_j приводит нас к выводу о том, что для рассматриваемого класса задач могут быть применены алгоритм 2 и доказанные для него оценки теоремы 2.3.7. Вообще говоря, уровни гладкости прямой и двойственной задач различны. Поэтому тут весьма важна также возможность адаптации разработанного метода к фактическому уровню гладкости задачи.

Теперь введём вариант неточного оракула (в смысле определения 2.2.1) для седловых задач вида (2.32) (это частный случай рассмотренного выше понятия неточной модели) и опишем связь с уже известным понятием (δ, L)-оракула для задач минимизации функционалов.

Определение 2.3.9. Пусть для фиксированного $\delta_u > 0$ (неконтролируемая ошибка) и для любого числа $\delta_c > 0$ (контролируемая ошибка) существует константа $L(\delta_c) \in (0, +\infty)$ такая, что при произвольных $x = (u_x, v_x), y = (u_y, v_y) \in Q$ возможно вычислить некоторые $\widetilde{G}(x, \delta_c, \delta_u), \widetilde{G}(y, \delta_c, \delta_u) \in E^*$ так, что верны неравенства

$$\langle \widetilde{G}(y,\delta_c,\delta_u) - \widetilde{G}(x,\delta_c,\delta_u), y - z\rangle \leqslant \frac{L(\delta_c)}{2}\left(\|y-x\|^2 + \|y-z\|^2\right) + \delta_c + \delta_u,$$

$$f(u_y, v_x) - f(u_x, v_y) \leqslant \langle \widetilde{G}(y,\delta_c,\delta_u), y - x\rangle + \delta_u$$

Тогда оператор $\widetilde{G}(x, \delta_c, \delta_u)$ будем называть *неточным оракулом* ((δ, L)-оракулом) для задачи нахождения седловой точки функционала f.

Замечание 2.3.10. Напомним (см. определение 2.2.1), что δ_c — контролируемая ошибка оракула. И наоборот, δ_u — неконтролируемая ошибка.

Пример 2.3.11 (*Седловые задачи и понятие (δ, L)-оракула в оптимизации*)**.** Предположим, что доступна величина

$$G_\delta(x) = \begin{pmatrix} g_{\delta,u}(u,v) \\ -g_{\delta,v}(u,v) \end{pmatrix}, x = (u,v) \in Q,$$

где $(f_{\delta,u}, g_{\delta,u})$ — (δ, L)-оракул f относительно u, а $-(f_{\delta,v}, \widetilde{g}_{\delta,v})$ есть (δ, L)-оракул для $(-f)$ относительно v, см. (2.26). Пусть $\|x\| = \|(u,v)\| :=$

$\sqrt{\|u\|^2 + \|v\|^2}$. Аналогично примеру 2.2.6 для произвольного $x = (u_x, v_x), y = (u_y, v_y), z = (u_z, v_z) \in Q$:

$$\langle g_{\delta,u}(u_y, v_y) - g_{\delta,u}(u_x, v_x), u_y - u_z\rangle \leqslant \frac{L}{2}(\|u_y - u_z\|^2 + \|u_y - u_x\|^2) + 3\delta,$$
$$\langle -g_{\delta,v}(u_y, v_y) + g_{\delta,v}(u_x, v_x), v_y - v_z\rangle \leqslant \frac{L}{2}(\|v_y - v_z\|^2 + \|v_y - v_x\|^2) + 3\delta,$$

Далее, из неравенств

$$f(u_y, v_y) - f(u_x, v_y) \leqslant \langle g_{\delta,u}(y), u_y - u_x\rangle + \delta,$$
$$f(u_y, v_x) - f(u_y, v_y) \leqslant \langle -g_{\delta,v}(y), v_y - v_x\rangle + \delta$$

имеем

$$f(u_y, v_x) - f(u_x, v_y) \leqslant \langle G_\delta(y), y - x\rangle + 2\delta.$$

Итак, $G_\delta(y) = \widetilde{G}(y, \delta_c, \delta_u)$ удовлетворяет определению 2.3.9 при $\delta_u = 6\delta$, $\delta_c = 0$ и $L(\delta_c) = L$.

В условиях примера 2.3.11 аналогично доказательству теоремы 2.3.7 можно проверить, что достаточно сделать

$$O\left(\inf_{\nu\in[0,1]}\left(\frac{L(\delta_c)}{\varepsilon}\right)\cdot \max_{x\in Q} V(x, w_0)\right)$$

итераций алгоритма 2, чтобы найти пару $(\widehat{u}, \widehat{v})$ (приближение искомой седловой точки), для которой верна

$$\max_{v\in Q_2} f(\widehat{u}, v) - \min_{u\in Q_1} f(u, \widehat{v}) \leqslant \varepsilon + O(\delta_u + \delta_c).$$

Таким образом, возможно выделять достаточно широкие классы седловых задач (по аналогии с описанными примерами в разделе 1.4 для задач минимизации), к которым можно применять предложенную в настоящей главе методику.

2.4 Некоторые вычислительные эксперименты по разработанным методам для вариационных неравенств и седловых задач

Ранее в данной главе работы предложен адаптивный метод для вариационных неравенств и седловых задач в модельной общности (алгоритм 1). Для специально подобранной оптимизационной модели превращается в универсальный метод (алгоритм 2) для вариационных неравенств с ν-гёльдеровыми монотонными операторами ($\nu \in [0;1]$). Универсальный метод позволяет реализовать адаптивную настройку работы метода на параметр гладкости оператора ν с гарантированным сохранением оптимальных оценок скорости сходимости при $\nu = 0$ и $\nu = 1$. Помимо приложения к универсальному методу модельная общность позволяет обосновать для некоторых типов седловых задач с негладкими целевыми функционалами (композитные задачи, задачи с функционалами вида максимума конечного набора гладких функционалов) оценки скорости сходимости, свойственные для гладкого случая. В этом пункте мы рассмотрим несколько примеров вычислительных экспериментов со случайно сгенерированными данными для негладких седловых задач (не композитный случай), для которых можно экспериментально за счёт адаптивного выбора шага наблюдать скорость сходимости, существенно лучшую по сравнению с известной оптимальной на соответствующем классе задач теоретической оценкой $O\left(\frac{1}{\varepsilon^2}\right)$. Отметим, что данные для лагранжевых седловых задач сгенерированы случайно и положение искомой точки-решения не очевидно. Поэтому разумно использовать именно теоретические оценки для понимания качества найденного приближения. Задачи и способ генерирования параметров для проведения экспериментов подобраны автором работы, программные коды на CPython 3.7 для результатов разделов 2.4.1, 2.4.3 и 2.4.4 подготовлены с помощью М. С. Алкусы и А. Н. Степанова.

2.4.1 Пример седловой задачи с негладким выпукло-вогнутым функционалом

Начнём с примера седловой задачи с негладким функционалом.

Пусть $x = (u, v) \in \mathbb{R}^p \times \mathbb{R}^q := \mathbb{R}^n$; $n = p + q$. Рассмотрим следующую

негладкую выпукло-вогнутую задачу седловой точки

$$\min_{u\in Q_1}\max_{v\in Q_2}\{f(u,v) := \|u-\alpha\|_2 + \langle Au-b, v\rangle - \|v-\beta\|_2\}, \tag{2.38}$$

где $Q_1 \times Q_2 \subset \mathbb{R}^p \times \mathbb{R}^q := \mathbb{R}^n$, $\alpha \in \mathbb{R}^p$, $\beta \in \mathbb{R}^q$ — фиксированные точки, $b \in \mathbb{R}^q$ и матрица A имеет размер $q \times p$. Ввиду негладкости f оператор G соответствующего вариационного неравенства

$$G(x) = \begin{pmatrix} \nabla_u f(u,v) \\ -\nabla_v f(u,v) \end{pmatrix},\ x = (u,v) \in Q := Q_1 \times Q_2.$$

гёльдеров при $\nu = 0$.

Пусть $p = 1000$, $q = 500$, т.е. $n = 1500$. Выберем значения параметров $\alpha_i = 0.1/\sqrt{1000}$, $(i = 1,...,p)$ и $\beta_j = -0.1/\sqrt{500}$, $(j = 1,...,q)$, а также $Q_1 = \{u \in \mathbb{R}^p : \|u\|_2 \leqslant 1\}$ и $Q_2 = \{v \in \mathbb{R}^q : \|v\|_2 \leqslant 1\}$. Будем рассматривать $\varepsilon \in \{1/(2i+2), i = 0,1,2,3,5,7,9,11,13,15,18,23,31\}$,

$$L_0 = \frac{\|G(1,0,...,0) - G(0,1,0,0,...,0)\|}{\sqrt{2}} \text{ и } x^0 = \mathbf{0} \in \mathbb{R}^n.$$

Элементы матрицы A взяты с обычным нормальным распределением. Координаты вектора b — случайные целые числа, взятые с дискретным равномерным распределением в отрезке $[-10, 10]$. Проведены 10 экспериментов со случайным выбором координат, результаты которых усреднены для соответствующего количества итераций.

Результаты работы алгоритма 2 для задачи (2.38) представлены на рисунках 2.4.1 и 2.4.2 ниже. Эти результаты демонстрируют количество итераций, произведенных для достижения ε-решения задачи (2.38) и время выполнения алгоритма в секундах для вышеупомянутого разного значения точности ε.

Как известно [36], для вариационного неравенства с негладким оператором теоретическая оценка сложности итерации $O\left(\frac{1}{\varepsilon^2}\right)$ является оптимальной. Однако экспериментально из наклона линии на рис. 2.4.1 видно, что наблюдается сходимость $O\left(\frac{1}{\sqrt[3]{\varepsilon}}\right)$, что возможно ввиду адаптивности.

2.4.2 Один пример вариационного неравенства, для которого предложенный метод сходится существенно быстрее теоретических оценок

Рассмотрим вариационное неравенство с оператором $G : \mathbb{R}^N \to \mathbb{R}^N$ в форме

$$G(x_1, x_2, \ldots, x_N) =$$

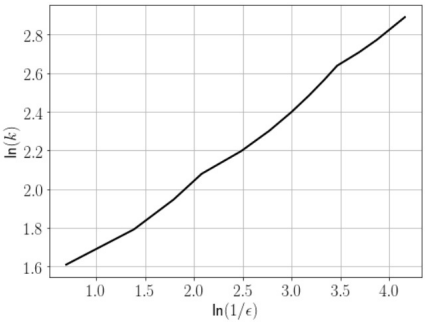

Рис. 2.4.1. Результаты для рассматриваемой седловой задачи.

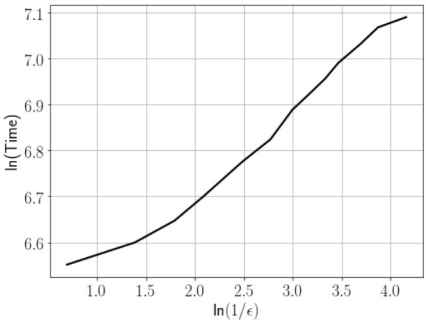

Рис. 2.4.2. Результаты для рассматриваемой седловой задачи.

$$= \left(\exp\left(x_1 + \frac{x_2}{\exp(3)}\right), \exp\left(x_2 + \frac{x_3}{\exp(3)}\right), \ldots, \exp\left(x_N + \frac{x_1}{\exp(3)}\right)\right). \tag{2.39}$$

В качестве допустимого множества Q выберем единичный шар с центром в нуле

$$Q = \left\{x = (x_1, x_2, \ldots, x_N) \mid x_1^2 + x_2^2 + \ldots + x_N^2 \leqslant 1\right\}$$

Норму выберем стандартной евклидовой $\mathbb{R}^N$:

$$\|x\| := \sqrt{x_1^2 + x_2^2 + \ldots + x_N^2}.$$

Выберем $x = (x_1, x_2, \ldots, x_N)$ и $y = (y_1, y_2, \ldots, y_N)$ — два вектора из Q. Очевидно, что оператор G не потенциален:

$$\frac{\partial}{\partial x_2}\left(\exp\left(x_1 + \frac{x_2}{\exp(3)}\right)\right) = \frac{1}{\exp(3)}\exp\left(x_1 + \frac{x_2}{\exp(3)}\right),$$

$$\frac{\partial}{\partial x_1}\left(\exp\left(x_2 + \frac{x_3}{\exp(3)}\right)\right) = 0.$$

Покажем, что оператор G удовлетворяет условию Липшица и монотонен на Q. По теореме о среднем значении для произвольных $i, j = \overline{1, N}$ имеем:

$$\exp\left(x_i + \frac{x_j}{\exp(3)}\right) - \exp\left(y_i + \frac{y_j}{\exp(3)}\right) =$$

$$= \exp\left(x_i + \frac{x_j}{\exp(3)}\right) - \exp\left(y_i + \frac{x_j}{\exp(3)}\right) +$$

$$+\exp\left(y_i+\frac{x_j}{\exp(3)}\right)-\exp\left(y_i+\frac{y_j}{\exp(3)}\right)=\exp\left(\alpha_i+\frac{x_j}{\exp(3)}\right)(x_i-y_i)+$$
$$+\frac{1}{\exp(3)}\exp\left(y_i+\frac{\gamma_j}{\exp(3)}\right)(x_j-y_j) \tag{2.40}$$

для некоторых α_i и γ_j : $|\alpha_i|\leqslant 1$ и $|\gamma_j|\leqslant 1$ (α_i и γ_j лежит между x_i, y_i и x_j, y_j соответственно). Понятно, что

$$\left|\alpha_i+\frac{x_j}{\exp(3)}\right|\leqslant\sqrt{1+\left(\frac{1}{\exp(3)}\right)^2}\sqrt{\alpha_i^2+x_j^2}<\sqrt{2},$$

а также $\left|y_i+\frac{\gamma_j}{\exp(3)}\right|<\sqrt{2}$. Следовательно, (2.40) означает, что

$$\left|\exp\left(x_i+\frac{x_j}{\exp(3)}\right)-\exp\left(y_i+\frac{y_j}{\exp(3)}\right)\right|<\exp(\sqrt{2})|x_i-y_i|+$$
$$+\exp(\sqrt{2}-3)|x_j-y_j|<\exp(\sqrt{2})\left(|x_i-y_i|+|x_j-y_j|\right).$$

Далее, с учетом неравенства $(a+b)^2\leqslant 2(a^2+b^2)$ имеем

$$\|G(x)-G(y)\|^2<4\cdot\exp(2\sqrt{2})\|x-y\|^2\quad\forall x,y\in Q,$$
$$\|G(x)-G(y)\|<2\cdot\exp(\sqrt{2})\|x-y\|\quad\forall x,y\in Q,$$

т.е. оператор G удовлетворяет свойству Липшица с константой $L=2\cdot\exp(\sqrt{2})$. Далее, (2.40) означает, что для любого $x,y\in Q$

$$\langle G(x)-G(y),x-y\rangle=$$
$$=\sum_{k=1}^{N}c_k(x_k-y_k)^2+\sum_{k=1}^{N-1}d_k(x_k-y_k)(x_{k+1}-y_{k+1})+d_N(x_N-y_N)(x_1-y_1),$$

где

$$c_k>\exp(-\sqrt{2}),\ d_k<\exp(\sqrt{2}-3)<\exp(-\sqrt{2})\quad\forall k=\overline{1,N}.$$

Взяв в расчет неравенство $2ab\leqslant a^2+b^2$, имеем

$$\langle G(x)-G(y),x-y\rangle>\exp(-\sqrt{2})\sum_{k=1}^{N}(x_k-y_k)^2-\exp(-\sqrt{2})\sum_{k=1}^{N}(x_k-y_k)^2=0,$$

таким образом оператор g является монотонным.

Замечание 2.4.1. Предположим, что на каждой итерации алгоритма 2 константа L_k делится не на 2, а на некоторую произвольную константу $a > 1$. Число обращений к оракулу (подпрограмме вычисления оператора) на каждой итерации k равно $2i_k$. В тоже время $L_k = a^{i_k-2}L_{k-1}$, поэтому $i_k = 2+\log_a \frac{L_k}{L_{k-1}}$. Пусть на $(i+1)$-й итерации это число составляло m_{i+1}. Тогда из-за деления на 2 в алгоритме 2 для всех $i = 0, 1, 2, ...$ имеем:

$$L_{i+1} = \frac{1}{2}2^{m_{i+1}-1}L_i = 2^{m_{i+1}-2}L_i,$$

откуда с учетом двух обращений к оракулу на каждой итерации общее количество обращений к подпрограмме для нахождения оператора G не будет превышать

$$\sum_{j=0}^{k-1} i_j = 4k + 2\sum_{i=0}^{k-1} \log_a \frac{L_{j+1}}{L_j} < 4k + 2\log_a aL - 2\log_a(L_0),$$

поскольку $L_k \leqslant aL$. Таким образом, количество обращений к оракулу для алгоритма 2 не превышает

$$4 \inf_{\nu\in[0,1]} \left(\frac{2\cdot L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}} \cdot \max_{u\in Q} V(u, z_0) +$$

$$+2 \inf_{\nu\in[0,1]} \log_a 2 \left(\left(\frac{2}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}\right) - 2\log_a(L_0).$$

Ниже приведены результаты работы метода для N-мерного оператора (2.39). Начальная точка $x_0 = (0.1, 0.1, \ldots, 0.1)$, и $x_0 = \frac{x_0}{||x_0||}$ при $||x_0|| > 1$. В таблице 1 приведены результаты работы алгоритма 1, в таблице 2 — результат работы модифицированного варианта этого метода, в котором на каждой итерации постоянная L_k делится не на 2, а на 16. Следует подчеркнуть, что увеличение размерности существенно не снижает скорость работы алгоритма, что подтверждается таблицами ниже.

2.4.3 Лагранжева седловая задача для задачи Ферма–Торричелли–Штейнера

Рассмотрим пример вариационного неравенства с ограниченным оператором (такой оператор гёльдеров при $\nu = 0$). Оказывается, что в этом случае алгоритм 2

Таблица 2.1. Результаты для стандартной версии алгоритма 2.

dim		ε							
		10^{-1}	$5 \cdot 10^{-2}$	10^{-2}	$5 \cdot 10^{-3}$	10^{-3}	$5 \cdot 10^{-4}$	10^{-4}	$5 \cdot 10^{-5}$
10^3	N	5	6	8	9	11	12	15	16
	время, с	0.031	0.038	0.049	0.055	0.068	0.073	0.092	0.096
$2 \cdot 10^3$	N	5	6	8	9	11	12	15	16
	время, с	0.062	0.071	0.095	0.109	0.141	0.152	0.19	0.2
10^4	N	5	6	8	9	11	12	15	16
	время, с	0.322	0.372	0.504	0.53	0.624	0.692	0.854	0.952
$5 \cdot 10^4$	N	5	6	8	9	11	12	15	16
	время, с	1.466	1.807	2.271	2.616	3.256	3.515	4.433	4.73
10^5	N	5	6	8	9	11	12	15	16
	время, с	2.787	3.495	4.314	5.085	5.969	6.578	8.232	9.016

Таблица 2.2. Результаты для модифицированной версии алгоритма 2, связанной с делением на 16.

dim		ε							
		10^{-1}	$5 \cdot 10^{-2}$	10^{-2}	$5 \cdot 10^{-3}$	10^{-3}	$5 \cdot 10^{-4}$	10^{-4}	$5 \cdot 10^{-5}$
10^3	N	2	2	3	3	3	4	4	4
	время, с	0.013	0.013	0.019	0.021	0.02	0.024	0.027	0.024
$2 \cdot 10^3$	N	2	2	3	3	3	4	4	4
	время, с	0.024	0.024	0.036	0.036	0.039	0.054	0.052	0.161
10^4	N	2	2	3	3	3	4	4	4
	время, с	0.121	0.123	0.178	0.239	0.195	0.265	0.258	0.238
$5 \cdot 10^4$	N	2	2	3	3	3	4	4	4
	время, с	0.674	0.602	0.955	0.966	1.004	1.531	1.418	1.28
10^5	N	2	3	3	3	3	4	4	4
	время, с	1.219	1.221	1.967	1.921	1.95	2.527	2.594	2.469

может работать со скоростью, превосходящей оптимальные теоретические оценки. Покажем это на примере экспериментов. Рассматриваемый пример естественно связан с аналогом известной задачи Ферма–Торричелли–Штейнера, но с некоторыми негладкими функциональными ограничениями.

Для заданного набора N точек $A_k \in \mathbb{R}^n, k = 1, ..., N$ рассмотрим задачу выпуклого программирования

$$\min_{x \in Q} \left\{ f(x) := \sum_{k=1}^{N} \|x - A_k\|_2 \;\middle|\; \varphi_p(x) := \sum_{i=1}^{n} \alpha_{pi}|x_i| - 1 \leqslant 0 \ , \ p = 1, ..., m \right\},$$

где Q — выпуклый компакт, α_{pi} взяты со стандартным нормальным распределением, затем при $\alpha_{pi} < 0$ выбираем $\alpha_{pi} = 0$.

Соответствующая лагранжева седловая задача имеет вид

$$\min_{x \in Q} \max_{\overrightarrow{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_m)^T \in \mathbb{R}^m_+} \Phi(x, \lambda) := f(x) + \sum_{p=1}^{m} \lambda_p \varphi_p(x),$$

Эта задача эквивалентна вариационному неравенству с монотонным ограниченным оператором

$$G(x, \lambda) = \begin{pmatrix} \nabla f(x) + \sum\limits_{p=1}^{m} \lambda_p \nabla \varphi_p(x), \\ (-\varphi_1(x), -\varphi_2(x), \ldots, -\varphi_m(x))^T \end{pmatrix}.$$

Если выполнены условия Слейтера, то значения $\overrightarrow{\lambda}$ можно ограничить, и допустимым множеством задачи будет компакт. Поэтому предположим, что существует (потенциально очень большая) оценка для оптимального множителя Лагранжа $\overrightarrow{\lambda}^*$, которая позволяет компактифицировать допустимое множество для пары $(x, \overrightarrow{\lambda})$ и будет евклидовым шаром некоторого радиуса.

Запустим алгоритм 2 для различных значений n, m и N со стандартной евклидовой прокс-структурой и начальной точкой $(x^0, \overrightarrow{\lambda}^0) = \frac{1}{\sqrt{m+n}}(1, 1, \ldots, 1) \in \mathbb{R}^{n+m}$. Выполнение критерия остановки алгоритма 2 гарантирует достижение для приближенного решения $\left(\widetilde{x}, \widetilde{\overrightarrow{\lambda}}\right)$ следующего неравенства:

$$\max_{\overrightarrow{\lambda} \in \Omega_\lambda} \Phi\left(\widetilde{x}, \overrightarrow{\lambda}\right) - \min_{x \in Q} \Phi\left(x, \widetilde{\overrightarrow{\lambda}}\right) \leqslant \varepsilon, \text{то есть}$$

$$\max_{\overrightarrow{\lambda}\in\Omega_\lambda}\left\{f(\widetilde{x})+\sum_{p=1}^{m}\lambda_p\varphi_p(\widetilde{x})\right\}-\min_{x\in Q}\left\{f(x)+\sum_{p=1}^{m}\widetilde{\lambda_p}\varphi_k(x)\right\}\leqslant\varepsilon,$$

откуда $\forall\overrightarrow{\lambda}\in\Omega_\lambda$, и точного решения x_* задачи выпуклого программирования ($\widetilde{\lambda_p}\geqslant 0$, $\varphi_p(x_*)\leqslant 0$)

$$\varepsilon\geqslant f(\widetilde{x})+\sum_{p=1}^{m}\lambda_p\varphi_p(\widetilde{x})-f(x_*)-\sum_{p=1}^{m}\widetilde{\lambda_p}\varphi_p(x_*)\geqslant$$

$$\geqslant f(\widetilde{x})-f(x_*)+\sum_{p=1}^{m}\lambda_p\varphi_p(\widetilde{x}).$$

Это означает, что $f(\widetilde{x})-f(x_*)\leqslant\varepsilon$, причем $\varphi_p(\widetilde{x})\leqslant\frac{\varepsilon}{\lambda_p}$ или $\varphi_p(\widetilde{x})\leqslant 0$. Иными словами, выполнение критерия остановки метода гарантирует достижение нужного качество решения по функции задачи выпуклого программирования.

Точки A_k, $k=1,...,N$ выбираются случайным образом со стандартным нормальным распределением. Проведены 10 экспериментов со случайным выбором точек, результаты которых усреднены для соответствующего количества итераций. Результаты работы алгоритма 2 при указанных данных представлены в таблице 2.3 и на рисунках 2.4.3 и 2.4.4. Эти результаты демонстрируют количество итераций алгоритма 2 для достижения ε-решения поставленной задачи, время выполнения алгоритма указано в секундах для различных значений желаемой точности $\varepsilon\in\{1/2^i, i=1,2,3,4,5,6\}$.

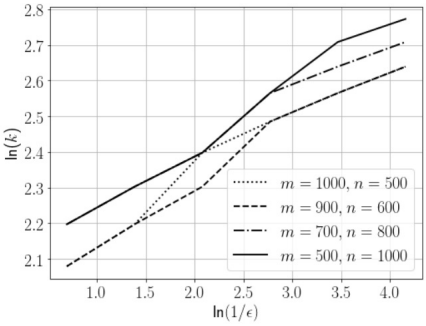

Рис. 2.4.3. Результаты выполнения алгоритма.

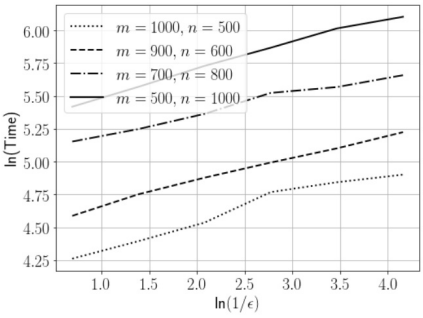

Рис. 2.4.4. Результаты выполнения алгоритма.

Таблица 2.3. Результаты работы универсального метода для седловых задач для задачи Ферма–Торричелли–Штейнера с различными значениями n, m и $N = 50$.

	Итерации	Время, с	Итерации	Время, с
$1/\varepsilon$	$n = 1000, m = 500$		$n = 800, m = 700$	
2	9	172.900	9	113.137
4	10	182.876	10	122.471
8	12	241.654	12	152.083
16	13	247.598	13	161.263
32	14	254.012	14	174.860
64	16	306.564	15	187.216
$1/\varepsilon$	$n = 600, m = 900$		$n = 500, m = 1000$	
2	8	60.559	8	45.009
4	10	77.569	9	48.692
8	11	84.827	10	52.689
16	12	93.686	12	66.830
32	14	108.939	13	71.409
64	15	115.100	14	76.283

2.4.4 Лагранжева седловая задача для задачи о наименьшем покрывающем шаре (минимизации максимума расстояний до фиксированного набора точек)

Рассмотрим пример вариационного неравенства с негладким и несильно монотонным оператором, то есть $\nu = 0$, для которого предложенный универсальный метод благодаря его адаптивности к уровню гладкости задачи на практике работает с итерационной сложностью, намного меньшей, чем предсказывает теория. Этот пример естественно связан с аналогом известной *задачи о наименьшем покрывающем круге* с некоторыми негладкими функциональными ограничениями.

Для заданного набора N точек $A_k \in \mathbb{R}^n, k = 1, ..., N$ нам нужно решить задачу выпуклого программирования

$$\min_{x \in Q} \left\{ f(x) := \max_{1 \leqslant k \leqslant N} \|x - A_k\|_2 \;\middle|\; \varphi_p(x) := \sum_{i=1}^{n} \alpha_{pi} |x_i| - 1 \leqslant 0 \right\}$$

$$(p = 1, \ldots, m),$$

где Q — выпуклый компакт, неотрицательные множители α_{pi} взяты

случайно со стандартным нормальным распределения. Соответствующая лагранжева седловая задача имеет вид

$$\min_{x\in Q}\max_{\overrightarrow{\lambda}=(\lambda_1,\lambda_2,\ldots,\lambda_m)^T\in\mathbb{R}^m_+}\Phi(x,\lambda):=f(x)+\sum_{p=1}^{m}\lambda_p\varphi_p(x),$$

Эта задача эквивалентна вариационному неравенству с монотонным негладким оператором

$$G(x,\lambda)=\begin{pmatrix}\nabla f(x)+\sum\limits_{p=1}^{m}\lambda_p\nabla\varphi_p(x),\\(-\varphi_1(x),-\varphi_2(x),\ldots,-\varphi_m(x))^T\end{pmatrix}.$$

Аналогично предыдущему пункту 2.4.3 допустим, что существует верхняя оценка для оптимального множителя Лагранжа $\overrightarrow{\lambda}^*$, которая позволяет компактифицировать допустимое множество для пары $(x,\overrightarrow{\lambda})$, чтобы она была евклидовым шаром некоторого радиуса.

Был выполнен запуск алгоритма 2 для различных значений n и m со стандартной евклидовой прокс-структурой и начальной точкой $(x^0,\overrightarrow{\lambda}^0)=\frac{1}{\sqrt{m+n}}\mathbf{1}\in\mathbb{R}^{n+m}$, где $\mathbf{1}$ — вектор всех единиц. Точки A_k, $k=1,\ldots,N$ выбираются случайным образом со стандартным нормального распределением. Проведены 5 экспериментов со случайным выбором координат этих точек. Результаты работы алгоритма 2 для этих экспериментов усреднены и представлены в таблице 2.4. Эти результаты демонстрируют количество итераций алгоритма 2 для теоретически гарантированного достижения ε-решения рассматриваемой задачи выпуклого программирования, а также время выполнения алгоритма в секундах с различными значениями точности $\varepsilon\in\{1/2^i, i=1,2,3,4,5,6\}$. Напомним, что в предыдущем пункте показано, что выполнение критерия остановки метода гарантирует достижение нужного качество решения по функции задачи выпуклого программирования.

Таблица 2.4. Результаты работы универсального метода с различными значениями n, m и $N = 50$.

	Итерации	Время, с	Итерации	Время, с
$1/\varepsilon$	$n = 1000, m = 500$		$n = 800, m = 700$	
2	9	21.351	9	16.904
4	10	24.880	10	18.977
8	11	27.718	11	18.977
16	13	32.817	12	23.434
32	14	35.265	14	26.438
64	15	38.464	15	26.438
$1/\varepsilon$	$n = 600, m = 900$		$n = 500, m = 1000$	
2	8	10.770	8	8.734
4	9	12.606	9	9.986
8	10	12.606	10	10.434
16	12	16.396	11	12.583
32	13	18.017	12	13.903
64	14	19.381	14	15.438

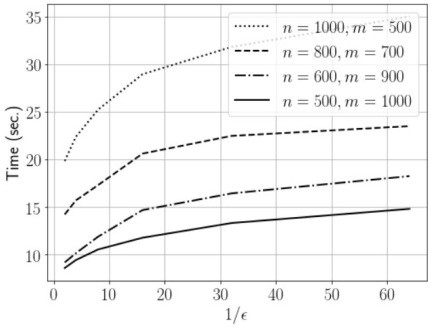

Рис. 2.4.5. Результаты работы универсального метода для задачи минимизации покрывающего шара с разными значениями m, n и $N = 50$.

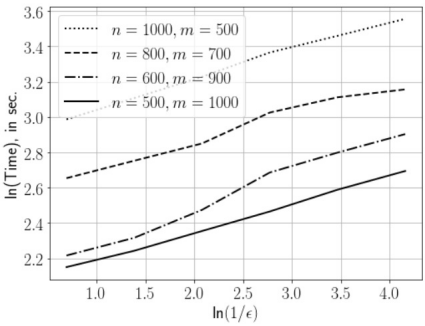

Рис. 2.4.6. Результаты работы универсального метода для задачи минимизации покрывающего шара с разными значениями m, n и $N = 50$.

2.4.5 Билинейные матричные игры Далее, рассмотрим расчеты для задачи нахождения равновесия по Нэшу в матричной игре.

Рассмотрим седловую задачу

$$\min_{x \in \Delta_n} \max_{y \in \Delta_m} x^T A y, \tag{2.41}$$

где $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, $y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^m$, Δ_n — единичный симплекс в $\mathbb{R}^n$, то есть $\Delta_n = \{x \in \mathbb{R}^n \,|\, x \geq 0, \sum\limits_{i=1}^{n+m} x_i = 1\}$, Δ_m — единичный симплекс в $\mathbb{R}^m$, A это матрица выплат для игрока y, размер матрицы $m \times n$. Рассмотрим следующий оператор

$$G(u) = \begin{pmatrix} \nabla_x (x^T A y) \\ -\nabla_y (x^T A y) \end{pmatrix} = \begin{pmatrix} A^T y \\ -Ax \end{pmatrix}, \; u = (x, y) \in Q \equiv \Delta_n \times \Delta_m. \tag{2.42}$$

Оператор $G(u)$ из (3.38) монотонен на Q. Поэтому для такого оператора вариационное неравенство

$$\langle G(x), x_* - x \rangle \leqslant 0 \;\; \forall\, x \in Q.$$

имеет решение, совпадающее с решением седловой задачи (3.37).

Поэтому возможно использовать алгоритм 2. В экспериментах рассматривается энтропийная прокс-функция вида $d(x) = \sum\limits_{i=1}^{n+m} x_i \ln x_i$ и соответствующая ей дивергенция Брегмана (расхождение Кульбака–Лейбнера). Такой выбор естественен для рассматриваемой постановки [38].

Результаты выполнения алгоритма представлены в таблице 2.5. В диссертации Ю. Е. Нестерова была показана теоретическая оценка для таких задач, согласно которой для достижения точности ε необходимо не менее $\frac{c}{\varepsilon}$ шагов, где $c > 0$ — некоторая постоянная. Как видим, за счет адаптивного выбора шага предложенный нами метод работает быстрее.

2.5 Адаптивный метод для вариационных неравенств с липшицевым сильно монотонным оператором

Данный раздел посвящен адаптивному методу для вариационных неравенств с L-липшицевым сильно монотонным оператором с адаптивной настройкой на величину параметра L. Предлагаемый метод основан на

Таблица 2.5. Результаты работы алгоритма, $m = 2500,\ n = 2500$.

ε	Итерации	Время, с
$1/2$	4.0	1.7
$1/4$	5.0	2.9
$1/6$	7.7	5.6
$1/8$	9.3	6.8
$1/10$	10.6	8.0
$1/12$	11.9	8.9
$1/14$	13.3	10.6
$1/16$	14.8	12.1
$1/18$	16.2	13.2
$1/20$	16.8	14.0

подходах диссертации [38]. Доказано, что скорость сходимости предложенного метода линейная и экспериментально показано, что адаптивная настройка параметра L может существенно повышать качество найденного решения по сравнению с неадаптивным вариантом рассматриваемого метода [38].

Хорошо известно, что в задачах оптимизации замена условия выпуклости функционала на сильную выпуклость приводит к существенно лучшей скорости сходимости методов. Аналогичный эффект имеет место и для вариационных неравенств, если оператор обладает свойством сильной монотонности. В [38, 143] был предложен метод для вариационных неравенств с сильно монотонным липшицевым оператором. Этот метод представляет собой комбинацию двойственного экстраполяционного метода [142] и методики оценочных функций (см. раздел 2.2 из [39]).

В данном разделе мы опишем предложенный нами адаптивный аналог метода Ю. Е. Нестерова для вариационных неравенств с липшицевым и сильно монотонным оператором, реализация которого не требует знания никакой верхней оценки $\widehat{L} \geqslant L$ константы Липшица L оператора G.

Будем рассматривать задачу нахождения решения $x_* = x_*(Q)$ вариационного неравенства

$$\langle G(x_*), x_* - y\rangle \leqslant 0 \quad \forall y \in Q, \tag{2.43}$$

где $G : Q \to \mathbb{R}^n$ — сильно монотонный оператор с параметром $\mu > 0$:

$$\langle G(x) - G(y), x - y\rangle \geqslant \mu||x - y||^2 \quad \forall x, y \in Q, \tag{2.44}$$

Q — выпуклое замкнутое (уже не обязательно компактное) подмножество $\mathbb{R}^n$, $\langle\cdot,\cdot\rangle$ — скалярное произведение в $\mathbb{R}^n$,

$$\|x\| = \langle Bx, x\rangle^{1/2} \tag{2.45}$$

есть некоторая евклидова норма в $\mathbb{R}^n$, где $B : \mathbb{R}^n \to \mathbb{R}^n$ — фиксированный оператор $B = B^T > 0$. Будем полагать, что оператор G удовлетворяет условию Липшица:

$$||G(x) - G(y)||_* \leqslant L||x - y|| \ \forall x, y \in Q \tag{2.46}$$

для некоторой константы $L > 0$, $||s||_* = \langle s, B^{-1}s\rangle^{1/2}$.

Напомним некоторые вспомогательные оценки, понятия и результаты из п. 3.2 диссертации Ю. Е. Нестерова [38]. Отметим, что сильная монотонность G означает, что для решения x_* верны оценки при произвольном $y \in Q$:

$$\langle G(y), x_* - y\rangle + \frac{\mu}{2}||y - x_*||^2 \leqslant \langle G(x_*), x_* - y\rangle - \frac{\mu}{2}||y - x_*||^2 \leqslant 0. \tag{2.47}$$

Неравенства (2.47) приводят к идее рассматривать следующую меру близости для оценки качества найденного приближённого решения x ВН (2.43):

$$\rho(x) = \sup_{y\in Q}\left\{\langle G(y), x - y\rangle + \frac{\mu}{2}||y - x||^2\right\}. \tag{2.48}$$

Отметим основные свойства ρ из (2.48).

Теорема 2.5.1. *(Ю. Е. Нестеров, [38]) Функция ρ из* (2.48) *определена и сильно выпукла на Q с параметром μ. Более того, для всякого $x \in Q$ $\rho(x) \geqslant 0$ и $\rho(x) = 0 \Leftrightarrow x = x_*$.*

Пусть в ходе работы некоторого алгоритма образовалась последовательность $\{y_i\}_{i=0}^N \subset Q$ и $\{\lambda_i\}_{i=0}^N$ — некоторый набор положительных чисел. Тогда обозначим

$S_N = \sum\limits_{i=0}^N \lambda_i$ и $\widetilde{y}_N := \dfrac{1}{S_N}\sum\limits_{i=0}^N \lambda_i y_i$ — усредненный выход работы алгоритма.

Неравенства (2.47) приводят к идее ввести следующую функцию зазора для оценки качества найденного решения:

$$\Delta_N := \max_{x \in Q} \left\{ \sum_{i=0}^{N} \lambda_i \left[\langle G(y_i), y_i - x \rangle - \frac{\mu}{2} ||x - y_i||^2 \right] \right\}. \tag{2.49}$$

Лемма 2.5.2. *(Ю. Е. Нестеров, [38]) Справедливо неравенство:* $\rho(\widetilde{y}_N) \leqslant \frac{\Delta_N}{S_N}$.

Вслед за [38] обозначим

$$\varphi_y^\beta(x) := \langle G(y), y - x \rangle - \frac{\beta}{2} ||x - y||^2, \quad \Phi_k(x) := \sum_{i=0}^{k} \lambda_i \varphi_{y_i}^\mu(x)$$

для произвольного параметра $\beta > 0$, $k = 0, 1, 2, \ldots$, а также $x, y \in Q$. Ясно, что функция φ_y^β сильно вогнута с параметром β, а $\Phi_k(x)$ сильно вогнута с параметром μS_k. Заметим, что при этом $(k = 0, 1, 2, \ldots, N)$

$$\Delta_k = \max_{x \in Q} \Phi_k(x).$$

Напомним метод Ю. Е. Нестерова для ВН с липшицевым сильно монотонным оператором [38,143]. Опишем $(k+1)$-ю итерацию этого метода $(k = 0, 1, 2, \ldots)$.

Алгоритм 3 Метод для ВН с сильно монотонным оператором

$x_k := \arg\max_{x \in Q} \Phi_k(x)$, $y_{k+1} := \arg\max_{x \in Q} \varphi_{x_k}^L(x)$, $\lambda_{k+1} := \frac{\mu}{L} S_k$.

Выход: $\widetilde{y}_{k+1} := \frac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i y_i$.

Для приведённого выше метода (алгоритм 3) согласно теореме 3.2.3 из [38] в случае липшицева оператора G с константой L для числа обусловленности $\gamma = \frac{L}{\mu}$ и произвольного натурального k верны оценки:

$$\frac{\mu}{2} ||\widetilde{y}_k - x_*||^2 \leqslant \rho(\widetilde{y}_k) \leqslant \left[\rho(y_0) + \frac{\mu(\gamma^2 - 1)}{2} ||y_0 - x_*||^2 \right] \exp\left(-\frac{k}{\gamma + 1} \right) \leqslant \tag{2.50}$$

$$\leqslant \rho(y_0) \cdot \gamma^2 \cdot \exp\left(-\frac{k}{\gamma + 1} \right). \tag{2.51}$$

Замечание 2.5.3. В конце пункта 3.2.2 [38] было проведено сравнение алгоритма 3 со стандартным проекционным методом вида

$$x_0 = \overline{x} \in Q, \tag{2.52}$$

$$x_{k+1} = \pi_Q(x_k - \lambda B^{-1} G(x_k)), \quad k \geqslant 0, \tag{2.53}$$

где $\pi_Q(x)$ — евклидова проекция точки x на множество Q. Как отмечено в ([38], конец п. 3.2.2), у этого метода может быть медленная сходимость. В частности, при выборе оптимального шага $\lambda = \dfrac{\mu}{L^2}$ верна оценка:

$$\|x_{k+1} - x_*\|^2 \leqslant \|x_k - x_*\|^2 \cdot \exp\left\{-\frac{k}{\gamma^2}\right\}.$$

При больших значениях числа обусловленности $\gamma = \frac{L}{\mu}$ эта оценка может быть значительно хуже, чем (2.50)–(2.51). Также известно, что скорость сходимости (2.50)–(2.51) не может быть улучшена никаким черноящичным методом, применяемым к задаче (2.43)–(2.44) (см. замечание в конце п. 3.2.2 [38]). В то же время с точки зрения сложности реализации метод Ю. Е. Нестерова не будет значительно сложнее метода (2.52)–(2.53): на каждой итерации требуется вычислить две проекции на множество и два значения оператора вместо одной проекции и одного значения в методе (2.52)–(2.53).

Теперь перейдём к основным результатам работы и предложим адаптивный аналог метода Ю. Е. Нестерова для ВН (2.43)–(2.44). Положим изначально $\lambda_0 := 1$, y_0 — некоторое начальное приближение искомого решения и выберем некоторое $0 < \beta_0 \leqslant 2L$, где L — константа Липшица для оператора G, удовлетворяющего условию (2.46).

Замечание 2.5.4. Ввиду сильной монотонности оператора G для произвольных различных x и y из множества Q верно $G(x) \neq G(y)$. Поэтому выполнения условия $\beta_0 \leqslant 2L$ можно добиться, выбрав

$$\beta_0 := \frac{\|G(x) - G(y)\|_*}{\|x - y\|} \tag{2.54}$$

для некоторых фиксированных различных x и y из Q.

Опишем $(k+1)$-ую итерацию предлагаемого метода ($k = 0, 1, 2, \ldots$).

Алгоритм 4 Адаптивный метод для ВН с сильно монотонным оператором

1. $x_k := \arg\max\limits_{x \in Q} \Phi_k(x),\ \beta_{k+1} := \frac{\beta_k}{2}$.

2. $y_{k+1} := \arg\max\limits_{x \in Q} \varphi_{x_k}^{\beta_{k+1}}(x)$.

3. **Если** верно

$$||G(y_{k+1}) - G(x_k)||_* \leqslant \sqrt{\beta_{k+1}(\beta_{k+1} + \mu)} \cdot ||y_{k+1} - x_k||, \qquad (2.55)$$

то вычисляем $\lambda_{k+1} := \frac{\mu}{\beta_{k+1}} S_k$, увеличиваем k на 1 и переходим к следующей итерации (п. 1).
Иначе $\beta_{k+1} := 2 \cdot \beta_{k+1}$ и переходим к п. 2.
Выход: $\widetilde{y}_{k+1} := \frac{1}{S_{k+1}} \sum\limits_{i=0}^{k+1} \lambda_i y_i$.

.

Замечание 2.5.5. При $\beta_{k+1} \geqslant L$ критерий выхода из итерации (2.55) заведомо выполнен, т.к. $\sqrt{L(L+\mu)} > L$ при всяком $\mu > 0$. Поэтому после завершения итерации алгоритма 4 заведомо верно неравенство $\beta_{k+1} < 2L$. Таким образом, константа β_{k+1} не может неограниченно увеличится и максимальное её значение будет сопоставимо с L.

Замечание 2.5.6. Аналогично рассуждениям из ([152], стр. 391) оценим количество операций п. 2 алгоритма 4. Пусть на $(k+1)$-й итераций их было i_{k+1}. Тогда ввиду деления на 2 в п. 1 алгоритма 4 мы имеем:

$$\beta_{k+1} = \frac{1}{2} 2^{i_{k+1}-1} \beta_k = 2^{i_{k+1}-2} \beta_k,$$

откуда

$$\sum_{k=0}^{N-1} i_{k+1} = 2N + \sum_{k=0}^{N-1} \log_2 \frac{\beta_{k+1}}{\beta_k} < 2N + \log_2(2L) - \log_2(\beta_0).$$

Таким образом, за счёт повторения вычислений в п. 2 сложность работы предлагаемого алгоритма 4 по сравнению с алгоритмом 3 может увеличится не более чем в 2 раза с точностью до постоянного слагаемого, зависящего от β_0 и L. Это означает, что трудоёмкость предлага-

емого метода вполне сопоставима с трудоёмкостью исходного алгоритма 3. Однако при этом не требуется знания никакой константы $\widehat{L} \geqslant L$. Преимуществом также является возможное существенное увеличение скорости сходимости метода в конкретных задачах (см., например, таблицу 2.6 ниже).

Справедлива следующая

Теорема 2.5.7. *При выполнении алгоритма 4 для величин* Δ_k *из* (2.49) *верно неравенство* $\Delta_{k+1} \leqslant \Delta_k$ *для всякого целого неотрицательного* k.

Доказательство. Ясно, что $\Phi_{k+1}(x) = \Phi_k(x) + \lambda_{k+1}\varphi^{\mu}_{y_{k+1}}(x)$. Тогда

$$\Delta_{k+1} = \max_{x\in Q}\left\{\Phi_k(x) + \lambda_{k+1}\varphi^{\mu}_{y_{k+1}}(x)\right\} \leqslant$$

$$\leqslant \Delta_k + \max_{x\in Q}\left\{\langle\nabla\Phi_k(x_k), x - x_k\rangle - \frac{1}{2}\mu S_k||x - x_k||^2 + \lambda_{k+1}\varphi^{\mu}_{y_{k+1}}(x)\right\} \leqslant$$

$$\leqslant \Delta_k + \max_{x\in Q}\left\{-\frac{1}{2}\mu S_k||x - x_k||^2 + \right.$$

$$\left. + \lambda_{k+1}\left[\langle G(y_{k+1}), y_{k+1} - x\rangle - \frac{1}{2}\mu||x - y_{k+1}||^2\right]\right\}.$$

В силу выбора y_{k+1} из п. 2 алгоритма 4 для всякого $x \in Q$ имеем:

$$\langle -G(x_k) - \beta_{k+1}B(y_{k+1} - x_k), x - y_{k+1}\rangle \leqslant 0.$$

Далее, с учётом равенства

$$2\cdot\langle B(y_{k+1} - x_k), x - y_{k+1}\rangle = ||x - x_k||^2 - ||y_{k+1} - x_k||^2 - ||x - y_{k+1}||^2$$

получаем оценки

$$\langle G(y_{k+1}), y_{k+1} - x\rangle - \frac{\mu}{2}||x - y_{k+1}||^2 =$$

$$= \langle G(y_{k+1}) - G(x_k), y_{k+1} - x\rangle - \frac{\mu}{2}||x - y_{k+1}||^2 + \langle G(x_k), y_{k+1} - x\rangle \leqslant$$

$$\leqslant ||G(y_{k+1}) - G(x_k)||_*\cdot||y_{k+1} - x|| - \frac{\mu}{2}||x - y_{k+1}||^2 +$$

$$+\beta_{k+1}\langle B(y_{k+1} - x_k), x - y_{k+1}\rangle =$$

$$= ||G(y_{k+1}) - G(x_k)||_*\cdot||y_{k+1} - x|| - \frac{\mu}{2}||x - y_{k+1}||^2 +$$

$$+\frac{\beta_{k+1}}{2}||x-x_k||^2-\frac{\beta_{k+1}}{2}||y_{k+1}-x_k||^2-\frac{\beta_{k+1}}{2}||x-y_{k+1}||^2=$$

$$=||G(y_{k+1})-G(x_k)||_*\cdot||y_{k+1}-x||-\frac{\beta_{k+1}+\mu}{2}||x-y_{k+1}||^2+$$

$$+\frac{\beta_{k+1}}{2}||x-x_k||^2-\frac{\beta_{k+1}}{2}||y_{k+1}-x_k||^2\leqslant$$

$$\leqslant-\frac{1}{2\mu}\left(||G(y_{k+1})-G(x_k)||_*-||y_{k+1}-x)||\right)^2+$$

$$+\frac{||G(y_{k+1})-G(x_k)||_*^2}{2\cdot(\beta_{k+1}+\mu)}+\frac{\beta_{k+1}}{2}\left(||x-x_k||^2-||y_{k+1}-x_k||^2\right)\leqslant$$

$$\leqslant\frac{||G(y_{k+1})-G(x_k)||_*^2}{2\cdot(\beta_{k+1}+\mu)}+\frac{\beta_{k+1}}{2}\left(||x-x_k||^2-||y_{k+1}-x_k||^2\right).$$

Поэтому ввиду (2.55)

$$\langle G(y_{k+1}),y_{k+1}-x\rangle-\frac{\mu}{2}||x-y_{k+1}||^2\leqslant\frac{1}{2(\beta_{k+1}+\mu)}||G(y_{k+1})-G(x_k)||_*^2+$$

$$+\frac{\beta_{k+1}}{2}||x-x_k||^2-\frac{\beta_{k+1}}{2}||y_{k+1}-x_k||^2\leqslant\frac{\beta_{k+1}}{2}||x-x_k||^2,$$

откуда с учетом $\mu S_k=\lambda_{k+1}\beta_{k+1}$ получаем требуемое. □

Следствие 2.5.8. *При выполнении алгоритма 4 верно неравенство* $\rho(\widetilde{y}_k)\leqslant\Delta_0\exp\left(-\frac{k\mu}{\mu+\widehat{\beta}}\right)$ *для всякого натурального* k*, где* $\widehat{\beta}$ *определяется следующим образом:*

$$1-\frac{\mu}{\mu+\widehat{\beta}}=\sqrt[k]{\left(1-\frac{\mu}{\mu+\beta_1}\right)\left(1-\frac{\mu}{\mu+\beta_2}\right)\ldots\left(1-\frac{\mu}{\mu+\beta_k}\right)}.\qquad(2.56)$$

Доказательство. Действительно, $S_0=\lambda_0=1$ и для всякого $k=0,1,\ldots$ верно

$$S_{k+1}=S_k+\lambda_{k+1}=\left(1+\frac{\mu}{\beta_{k+1}}\right)S_k.$$

Далее, по лемме 2.5.2

$$\rho(\widetilde{y}_k)\leqslant\frac{\Delta_k}{S_k}\leqslant\Delta_0\cdot\frac{S_0}{S_1}\cdot\frac{S_1}{S_2}\cdot\ldots\cdot\frac{S_{k-1}}{S_k}=$$

$$=\Delta_0\left(1-\frac{\mu}{\mu+\beta_1}\right)\left(1-\frac{\mu}{\mu+\beta_2}\right)\ldots\left(1-\frac{\mu}{\mu+\beta_k}\right)=$$

$$= \Delta_0 \left(1 - \frac{\mu}{\mu + \widehat{\beta}}\right)^k = \Delta_0 \left(1 - \frac{1}{1 + \frac{\widehat{\beta}}{\mu}}\right)^k \leqslant \Delta_0 \exp\left(-\frac{k\mu}{\mu + \widehat{\beta}}\right),$$

что и требовалось. □

Из теорем 2.5.1 и 2.5.7, а также следствия 2.5.8 вытекает следующий результат, аналогичный теореме 3.2.3 из [38].

Теорема 2.5.9. *Пусть оператор G липшицев с константой $L > 0$ и сильно монотонен с параметром $\mu > 0$. Тогда при выполнении алгоритма 4 для $\gamma = \frac{L}{\mu}$ и всякого натурального k верны оценки:*

$$\frac{\mu}{2}||\widetilde{y}_k - x_*||^2 \leqslant \rho(\widetilde{y}_k) \leqslant \left[\rho(y_0) + \frac{\mu(\gamma^2 - 1)}{2}||y_0 - x_*||^2\right] \exp\left(-\frac{k}{1 + \frac{\widehat{\beta}}{\mu}}\right) \leqslant \tag{2.57}$$

$$\leqslant \rho(y_0) \cdot \gamma^2 \cdot \exp\left(-\frac{k}{1 + \frac{\widehat{\beta}}{\mu}}\right). \tag{2.58}$$

Отметим, что оценки (2.57)–(2.58) могут оказаться лучше (2.50)–(2.51) из [38], поскольку $\frac{\widehat{\beta}}{\mu}$ может оказаться меньше γ. Далее, это наглядно продемонстрировано на примере численного эксперимента для задачи (2.59).

Замечание 2.5.10. Рассмотрим модификацию алгоритма 4, которая исключает уменьшение константы β_{k+1} в ходе работы метода. Это даёт возможность сделать вывод о несущественном увеличении трудоёмкости по сравнению с методом Ю. Е. Нестерова (алгоритм 3). Изначально положим $\lambda_0 := 1$, y_0 — некоторое начальное приближение искомого решения и выберем некоторое $0 < \beta_0 \leqslant 2L$ (см. (2.54)), где L — константа Липшица для оператора G из (4.21). Опишем $(k+1)$-ю итерацию предлагаемой модификации алгоритма 4 ($k = 0, 1, 2, \ldots$).

Алгоритм 5 Модификация алгоритма 4

1. $x_k := \arg\max\limits_{x \in Q} \Phi_k(x),\ \beta_{k+1} := \beta_k$.

2. $y_{k+1} := \arg\max\limits_{x \in Q} \varphi_{x_k}^{\beta_{k+1}}(x)$.

3. **Если** верно

$$||G(y_{k+1}) - G(x_k)||_* \leqslant \sqrt{\beta_{k+1}(\beta_{k+1} + \mu)} \cdot ||y_{k+1} - x_k||,$$

то вычисляем $\lambda_{k+1} := \frac{\mu}{\beta_{k+1}} S_k$, увеличиваем k на 1 и переходим к следующей итерации (п. 1).
Иначе $\beta_{k+1} := 2 \cdot \beta_{k+1}$ и переходим к п. 2.
Выход: $\widetilde{y}_{k+1} := \frac{1}{S_{k+1}} \sum\limits_{i=0}^{k+1} \lambda_i y_i$.

.

Поскольку β_{k+1} может лишь увеличиваться, то по сравнению с алгоритмом 3 количество вычислений согласно п. 2 возрастёт лишь не более чем на

$$\left\lceil \log_2 \frac{2L}{\beta_0} \right\rceil.$$

Для демонстрации преимуществ алгоритмов 4 и 5 по сравнению с алгоритмом 3 были проведены вычислительные эксперименты для вариационного неравенства с оператором $G : \mathbb{R}^n \to \mathbb{R}^n$ $(n = 1000)$ вида

$$\begin{aligned} G(x_1, x_2, \ldots, x_n) = \\ = \left(\exp\left(x_1 + \frac{x_2}{10\exp(3)} \right), \exp\left(x_2 + \frac{x_3}{10\exp(3)} \right), \ldots, \right. \\ \left. \exp\left(x_n + \frac{x_1}{10\exp(3)} \right) \right). \end{aligned} \tag{2.59}$$

В качестве множества Q выберем единичный шар с центром в нуле

$$Q = \left\{ x = (x_1, x_2, \ldots, x_n) \mid x_1^2 + x_2^2 + \ldots + x_n^2 \leqslant 1 \right\}$$

для стандартной евклидовой нормы в $\mathbb{R}^n$ (т.е. здесь оператор B в (2.45) мы полагаем единичным): $\|x\| := \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$.

Пусть $x = (x_1, x_2, \ldots, x_n)$ и $y = (y_1, y_2, \ldots, y_n)$ — два вектора из Q. Очевидно, что оператор G не является потенциальным:

$$\frac{\partial}{\partial x_2}\left(\exp\left(x_1 + \frac{x_2}{10\exp(3)}\right)\right) = \frac{1}{10\exp(3)}\exp\left(x_1 + \frac{x_2}{10\exp(3)}\right),$$

$$\frac{\partial}{\partial x_1}\left(\exp\left(x_2 + \frac{x_3}{10\exp(3)}\right)\right) = 0.$$

Покажем, что оператор G удовлетворяет условию Липшица и сильно монотонен на Q. По теореме о среднем для произвольных $i, j = \overline{1,1000}$ имеем:

$$\begin{aligned}
&\exp\left(x_i + \frac{x_j}{10\exp(3)}\right) - \exp\left(y_i + \frac{y_j}{10\exp(3)}\right) = \\
&= \exp\left(x_i + \frac{x_j}{10\exp(3)}\right) - \exp\left(y_i + \frac{x_j}{10\exp(3)}\right) + \\
&+ \exp\left(y_i + \frac{x_j}{10\exp(3)}\right) - \exp\left(y_i + \frac{y_j}{10\exp(3)}\right) = \\
&= \exp\left(\alpha_i + \frac{x_j}{10\exp(3)}\right)(x_i - y_i) + \\
&+ \frac{1}{10\exp(3)}\exp\left(y_i + \frac{\gamma_j}{10\exp(3)}\right)(x_j - y_j)
\end{aligned} \tag{2.60}$$

для некоторых α_i и γ_j : $|\alpha_i| \leqslant 1$ и $|\gamma_j| \leqslant 1$ (α_i и γ_j лежат между x_i, y_i и x_j, y_j соответственно). Ясно, что

$$\left|\alpha_i + \frac{x_j}{10\exp(3)}\right| \leqslant \sqrt{1 + \left(\frac{1}{10\exp(3)}\right)^2}\sqrt{\alpha_i^2 + x_j^2} < \sqrt{2},$$

а также $\left|y_i + \dfrac{\gamma_j}{10\exp(3)}\right| < \sqrt{2}$. Поэтому (2.60) означает, что

$$\left|\exp\left(x_i + \frac{x_j}{10\exp(3)}\right) - \exp\left(y_i + \frac{y_j}{10\exp(3)}\right)\right| < \exp(\sqrt{2})|x_i - y_i| +$$

$$+ \exp(\sqrt{2} - 3)\frac{|x_j - y_j|}{10} < \exp(\sqrt{2})\left(|x_i - y_i| + \frac{|x_j - y_j|}{10}\right).$$

Далее, с учётом неравенства $(a+b)^2 \leqslant 2(a^2 + b^2)$ имеем

$$\|G(x) - G(y)\|^2 < 2\exp(2\sqrt{2})\left(1 + \frac{1}{100}\right)\|x - y\|^2 \quad \forall x, y \in Q,$$

$$\|G(x) - G(y)\| < \frac{\sqrt{202}\exp(\sqrt{2})}{10}\|x - y\| \quad \forall x, y \in Q,$$

т.е. оператор G удовлетворяет свойству Липшица с константой $L = \frac{\sqrt{202}}{10}\exp(\sqrt{2})$.

Далее, (2.60) означает, что для произвольных $x, y \in Q$

$$\langle G(x) - G(y), x - y\rangle =$$

$$= \sum_{k=1}^{n} c_k(x_k - y_k)^2 + \sum_{k=1}^{19} d_k(x_k - y_k)(x_{k+1} - y_{k+1}) + d_n(x_n - y_n)(x_1 - y_1),$$

где

$$c_k > \exp(-\sqrt{2}),\ d_k < \frac{\exp(\sqrt{2} - 3)}{10} < \frac{\exp(-\sqrt{2})}{10}$$

для всякого $k = \overline{1, 1000}$. Учитывая неравенство $2ab \leqslant a^2 + b^2$, получаем

$$\langle G(x) - G(y), x - y\rangle > \exp(-\sqrt{2})\sum_{k=1}^{n}(x_k - y_k)^2 - \frac{\exp(-\sqrt{2})}{10}\sum_{k=1}^{n}(x_k - y_k)^2 =$$

$$= \frac{9}{10}\exp(-\sqrt{2})\|x - y\|^2,$$

т.е. оператор G сильно монотонен с параметром $\mu = \frac{9}{10}\exp(-\sqrt{2})$.

Мы применили алгоритмы 3, 4 и 5 к вариационному неравенству для оператора G из (2.59) с параметрами $L = \frac{\sqrt{202}}{10}\exp(\sqrt{2})$, $\mu = \frac{9}{10}\exp(-\sqrt{2})$, начального приближения $y_0 = (0.2, 0.2, ..., 0.2) \in Q$ и в соответствии с (2.54)

$$\beta_0 = \frac{\|G(1, 0, 0, ..., 0) - G(0, 1, 0, ..., 0)\|}{\sqrt{2}} \tag{2.61}$$

для стандартной евклидовой нормы в $\mathbb{R}^n$.

Результаты сравнения работы алгоритмов 3 и 4, а также алгоритмов 3 и 5 представлены выше в сравнительных таблицах 2.6 и 2.7. В указанных таблицах N — количество итераций работы этих алгоритмов, и время работы алгоритмов указано в миллисекундах. Величине, определяющей качество найденного решения согласно оценке скорости сходимости для алгоритма 3 и времени работы соответствуют второй и третий столбец указанных таблиц. Величинам, определяющим качество

Таблица 2.6. "Сравнение результатов работы алгоритмов 3 и 4".

N	$\exp\left(\frac{-k}{1+\frac{L}{\mu}}\right)$	Время, мс	$\exp\left(\frac{-k}{1+\frac{\widehat{\beta}}{\mu}}\right)$	Время, мс	β_N	$\widehat{\beta}$
3	8.9742e-01	14	1.0033e-01	53	5.3618e-02	6.6678e-02
6	8.0536e-01	27	1.0066e-02	102	1.0724e-01	6.6678e-02
9	7.2274e-01	44	1.3020e-03	148	1.0724e-01	7.7596e-02
12	6.4860e-01	55	1.4873e-04	207	1.0724e-01	7.9114e-02
15	5.8207e-01	70	1.6997e-05	262	1.0724e-01	8.0042e-02
18	5.2236e-01	82	1.6952e-06	299	5.3618e-02	7.7596e-02
21	4.6878e-01	93	1.6936e-07	316	5.3618e-02	7.5906e-02
24	4.2069e-01	106	1.6939e-08	371	5.3618e-02	7.4669e-02
27	3.7753e-01	121	2.2071e-09	404	1.0724e-01	7.7596e-02
30	3.3881e-01	135	2.8891e-10	417	1.0724e-01	8.0042e-02
33	3.0405e-01	155	3.7948e-11	474	1.0724e-01	8.2117e-02
36	2.7286e-01	179	4.9975e-12	520	1.0724e-01	8.3899e-02
39	2.4487e-01	183	6.5947e-13	577	1.0724e-01	8.5445e-02
42	2.1975e-01	200	8.7167e-14	596	1.0724e-01	8.6799e-02
45	1.9721e-01	221	1.1537e-14	645	1.0724e-01	8.7995e-02

Таблица 2.7. "Сравнение результатов работы алгоритмов 3 и 5".

N	$\exp\left(\frac{-k}{1+\frac{L}{\mu}}\right)$	Время, мс	$\exp\left(\frac{-k}{1+\frac{\widehat{\beta}}{\mu}}\right)$	Время, мс	β_N	$\widehat{\beta}$
3	8.9742e-01	14	2.1981e-01	55	2.1447e-01	2.1447e-01
6	8.0536e-01	27	1.1924e-01	78	3.4316	3.9854e-01
9	7.2274e-01	44	8.9413e-02	123	3.4316	5.9679e-01
12	6.4860e-01	55	7.5746e-02	157	6.8632	7.9875e-01
15	5.8207e-01	70	6.5966e-02	171	6.8632	9.8846e-01
18	5.2236e-01	82	5.8287e-02	206	6.8632	1.1668
21	4.6878e-01	93	5.1942e-02	229	6.8632	1.3348
24	4.2069e-01	106	4.6541e-02	263	6.8632	1.4932
27	3.7753e-01	121	4.1855e-02	317	6.8632	1.6428
30	3.3881e-01	135	3.7740e-02	347	6.8632	1.7843
33	3.0405e-01	155	3.4094e-02	354	6.8632	1.9183
36	2.7286e-01	179	3.0845e-02	384	6.8632	2.0455
39	2.4487e-01	183	2.7937e-02	407	6.8632	2.1663
42	2.1975e-01	200	2.5326e-02	465	6.8632	2.2812
45	1.9721e-01	221	2.2975e-02	542	6.8632	2.3906

найденного решения согласно оценкам скорости и времени работы для алгоритмов 4 соответствуют четвёртый и пятый столбцы этих таблиц.

Как видим из таблицы 2.6, скорость сходимости для предлагаемого нами алгоритма 4 существенно выше скорости сходимости алгоритма 3. Это получается за счёт значительного уменьшения констант β_N на итерациях в ходе работы алгоритма, а также предлагаемого нами их усреднения в (2.56). Из таблицы 2.7 видим, что скорость сходимости для предлагаемого нами алгоритма 5 выше, чем для алгоритма 3, но уже не так существенно, как для алгоритма 4. При этом время работы алгоритма 5 меньше, чем время работы алгоритма 4. По сути, преимущество алгоритма 5 перед алгоритмом 3 для рассматриваемого примера определяется, прежде всего, возможностью выбора начальной константы β_0 (2.61) согласно предлагаемому нами способу в замечании 2 без использования какой-либо оценки $\widehat{L} \geqslant L$ константы Липшица L оператора G.

Заключительные замечания к главе 2

Выделим основные результаты данной главы:

– В качестве существенного обобщения недавно введённого для задач минимизации понятия (δ, L)-оракула О. Деволдера–Ф. Глинера–Ю. Е. Нестерова, предложено понятие (δ, L)-модели для вариационных неравенств, седловых задач.

– Для выделенного класса задач предложен аналог известного для вариационных неравенств и седловых задач проксимального зеркального метода (алгоритм 1). При этом рассмотрено специальное условие гладкости, а также предложен адаптивный критерий остановки метода. Адаптивность позволяет применять метод для задач с неизвестной константой L, а также может приводить к более скорому достижению желаемой точности решения по сравнению с теоретическими вычислительными гарантиями.

– Получена оценка скорости сходимости алгоритма 1, указывающая на его оптимальность с точки зрения теории нижних оракульных оценок (теорема 2.1.6 и замечание 2.1.7). При этом учитываются погрешности задания функционала ψ_δ, а также решения вспомогательных задач на итерациях метода (см. (2.12)). Доказано, что погрешности обоих типов не накапливаются в ходе работы метода.

– Введено понятие (δ, L)-модели для седловых задач (определение 2.3.1). Получена оценка скорости сходимости алгоритма 2 для седловых задач, допускающих (δ, L)-модель (теорема 2.3.3). При этом также показана возможность учёта погрешности задания модели седловой задачи, а также погрешности решения вспомогательных задач на итерациях метода.

– Обоснована возможность применения предложенного метода к смешанным вариационным неравенствам (пример 2.0.2) и композитным седловым задачам (пример 2.3.2), а также к недавно рассмотренному в [65] варианту задачи о распределении ресурсов (пример 2.1.3);

– Как частный случай, предложен универсальный метод для вариационных неравенств с гёльдеровыми монотонными операторами и для соответствующего класса выпукло-вогнутых седловых задач (алгоритм 2);

– Предложен адаптивный метод для вариационных неравенств с липшицевыми сильно монотонными операторами (здесь уже Q необязательно компактно).

Таким образом, рассмотренное в настоящей главе понятие (δ, L)-модели функции для абстрактных ВН, позволило обосновать применимость адаптивного варианта проксимального зеркального метода к достаточно широкому классу задач. В частности, методика настоящей работы применима к смешанным вариационным неравенствам [31, 69], а также композитным седловым задачам [84]. При этом была учтена возможность неточного задания функционала ψ, а также неточность решения вспомогательных задач проектирования на итерациях метода. Показано, что при этом сохраняются оптимальные с точки зрения нижних оракульных оценок скорости сходимости метода, а погрешности обоих типов не накапливаются в ходе итераций метода. Разработанный на базе общего подхода универсальный метод для вариационных неравенств позволяет реализовать адаптивную настройку работы метода на параметр гладкости оператора ν с гарантированным сохранением оптимальных вычислительных гарантий в случаях $\nu = 0$ и $\nu = 1$. Помимо приложения к универсальному методу модельная общность позволяет обосновать для некоторых типов седловых задач с негладкими целевыми функционалами (композитные задачи) оценки скорости сходимости, свойственные для гладкого случая. Приводятся примеры, иллюстрирующие возможность повышения качества работы предложенного метода

по сравнению с теоретическими оценками за счёт адаптивного выбора шага и адаптивного критерия остановки. Результаты настоящей главы опубликованы в [17, 54, 179].

Отметим некоторые вопросы, интересные для дальнейшей более детальной проработки. Хорошо известно, что в общем случае к седловым задачам с помощью метода множителей Лагранжа сводятся задачи выпуклого программирования. Допустим, что целевой функционал или функционал ограничения в задаче условной оптимизации не имеет липшицева градиента, но при этом такой функционал допускает (δ, L)-модель в произвольной точке. В таком случае ожидается, что соответствующая лагранжева седловая задача будет иметь (δ, L)-модель в смысле определения 2.3.1. Отметим, что примеры таких функционалов для безусловных задач приведены в [22, 168]. Представляет интерес экспериментально исследовать эффективность полученных в данной главе результатов к задачам условной оптимизации с целевыми функционалами вида [168]. Также можно более детально исследовать лагранжевы седловые задачи для задач выпуклого программирования с относительно гладкими целевыми функционалами [67, 157].

ГЛАВА 3

Адаптивные методы для оптимизационных задач, допускающих существование аналогов неточного оракула с двумя параметрами, соответствующих погрешностям

Введение

В предыдущей главе был введён аналог концепции неточного оракула для вариационных неравенств и седловых задач с рассмотрением абстрактной модельной общности. Для класса задач, допускающих такую оптимизационную модель и предложен адаптивный аналог проксимального зеркального метода. При этом обоснована возможность адаптивной настройки только на параметр гладкости L и не было предусмотрено адаптивной настройки параметра δ, соответствующего возможным погрешностям. То же самое можно сказать и о ранее известных градиентных методах для концепций неточного оракула и абстрактной неточной модели функции для задач минимизации функционала. В частности, в [22,168] рассматривались методы с адаптивной настройкой только лишь на константу гладкости L при сохранении соответствующих погрешностям параметров постоянными.

В настоящей главе предложены новые, более продвинутые варианты понятий неточного оракула и неточной модели оптимизируемой функции для задачи минимизации функционалов, вариационных неравенств и седловых задач. Отличительная особенность — наличие не одного, а двух соответствующих возможным погрешностям параметров (в частности, для гладких задач погрешности задания целевого функционала и его градиента или субградиента). При этом удалось в некотором смысле обосновать возможность избежать накопления значений одного из этих параметров погрешностей (задания градиента в гладком случае) в

теоретических оценках скорости сходимости методов, а также рассмотреть некоторые классы негладких задач, к которым применимы такие результаты. Для неускоренных методов введенные понятия позволили предложить методы, применимые и к популярным в последние годы относительно гладким оптимизационным задачам при наличии погрешностей задания целевой функции и градиента (в этом случае уже несколько затруднительно применять известные подходы работ О. Деволдера, Ф. Глинера и Ю. Е. Нестерова в области задач с неточными оракулами, поскольку дивергенция $V(y, x)$ и $\|y - x\|^2$ уже никак в общем случае не связаны).

Подчеркнём, что основная общая идея всех обсуждаемых в данной главе подходов — возможность адаптивной настройки метода не только на параметр гладкости, но и на величины параметров оптимизационной модели задачи, соответствующих погрешностям. Это позволяет уменьшить влияние погрешностей на итоговые оценки скорости сходимости, что проиллюстрировано некоторыми вычислительными экспериментами (см. пример 3.2.7, а также пункт 3.4.3). Точнее говоря, вводится новый аналог концепции (δ, L)-модели целевой функции для задач минимизации f, которая позволяет раздельно учитывать возможность неточного задания значения целевой функции, а также градиента (в общем случае абстрактной выпуклой функции модели). В частности, для стандартной функции-модели $\psi(y, x) = \langle \nabla f(x), y - x \rangle$ эта ситуация описывается некоторой модификации условий

$$f(x) + \langle \nabla f(x), y - x \rangle - \delta \leqslant f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 + \delta \tag{3.1}$$

для некоторых $L > 0$ и $\delta > 0$ с учетом отдельно погрешности задания f и ∇f. Если положить, что для произвольного $x \in Q$ верно

$$\left\| \nabla f(x) - \widetilde{\nabla} f(x) \right\|_* \leqslant \Delta, \;\; \Delta > 0 \tag{3.2}$$

для некоторого доступного приближенного значения $\widetilde{\nabla} f(x)$ градиента ∇f, то будет верно неравенств $\left| \left\langle \nabla f(x) - \widetilde{\nabla} f(x), y - x \right\rangle \right| \leqslant \Delta \|y - x\|$, то есть

$$f(y) \leqslant f(x) + \left\langle \widetilde{\nabla} f(x), y - x \right\rangle + \frac{L}{2} \|y - x\|^2 + \Delta \|y - x\|,$$

а также

$$f(y) \geqslant f(x) + \left\langle \widetilde{\nabla} f(x), y - x \right\rangle - \Delta \|y - x\|$$

для всяких $x, y \in Q$. Если кроме этого предположить, что доступно неточное значение целевой функции $f_\delta(x)$: $f_\delta(x) \leqslant f(x) \leqslant f_\delta(x) + \delta$ при $\delta > 0$, то возможно выписать следующий аналог (3.1):

$$f_\delta(x) + \left\langle \widetilde{\nabla} f(x), y - x \right\rangle - \Delta \|y - x\| \leqslant f(y) \leqslant$$

$$\leqslant f_\delta(x) + \left\langle \widetilde{\nabla} f(x), y - x \right\rangle + \frac{L}{2} \|y - x\|^2 + \Delta \|y - x\| + \delta \ \ \forall x, y \in Q$$

или же

$$\begin{aligned} f(x) + \left\langle \widetilde{\nabla} f(x), y - x \right\rangle - \delta - \Delta \|y - x\| \leqslant f(y) \leqslant \\ \leqslant f(x) + \left\langle \widetilde{\nabla} f(x), y - x \right\rangle + \frac{L}{2} \|y - x\|^2 + \Delta \|y - x\| + \delta \ \ \forall x, y \in Q. \end{aligned} \tag{3.3}$$

В настоящей главе будет рассмотрен (в модельной общности) следующий аналог неравенства (3.3) с параметрами $\delta, \gamma, \Delta \geqslant 0$:

$$\begin{aligned} f(x) + \left\langle \widetilde{\nabla} f(x), y - x \right\rangle - \delta - \gamma \|y - x\| \leqslant f(y) \leqslant \\ \leqslant f(x) + \left\langle \widetilde{\nabla} f(x), y - x \right\rangle + \frac{L}{2} \|y - x\|^2 + \Delta \|y - x\| + \delta \ \ \forall x, y \in Q. \end{aligned} \tag{3.4}$$

Смысл такого обобщения заключается в том, что возможны различные значения параметров γ и Δ в (3.4). Как показано далее, влияние параметра Δ на качество решения, возвращаемое методом, может быть уменьшено. Отметим ещё, что далее в разделе 3.3 подробно разобрано несколько примеров негладких задач, когда $\delta = \gamma = 0$ при $\Delta > 0$. Если положить $\gamma = 0$, то $\widetilde{\nabla} f(x)$ — δ-субградиент f в точке x и параметр $\Delta > 0$ может указывать в этом случае на скачки $\widetilde{\nabla} f(x)$ в точках негладкости f. Если положить $\delta = 0$, то при $\gamma > 0$ $\widetilde{\nabla} f(x)$ — так называемый аналитический γ-субградиент f ([131], Sect. 1.3). В итоге мы предлагаем максимально общую концепцию неточной модели целевой функции, которая описывает все указанные ситуации. Для функций, допускающих существование такой модели в любой запрошенной точке, предлагается адаптивный градиентный метод (алгоритм 6) и доказывается теорема об оценке скорости его сходимости (теорема 3.1.4).

Выделим наиболее важные результаты настоящей главы:

– В разделе 3.1 обобщено ранее предложенное понятие (δ, L)-модели целевой функции в запрошенной точке и введена концепция $(\delta, \gamma, \Delta, L)$-модели функции (определение 3.1.1). Предложен градиентный метод

(алгоритм 6) для задач выпуклой минимизации с адаптивным выбором шага и адаптивной настройкой на некоторые из параметров $(\delta, \gamma, \Delta, L)$-модели, получена оценка качества решения в зависимости от номера итерации.

– В разделе 3.2 для более узкого класса оптимизационных задач, допускающих существование (δ, Δ, L)-модели в произвольной запрошенной точки, предложен аналог быстрого градиентного метода (алгоритм 7), получена оценка качества решения в зависимости от номера итерации.

– В разделе 3.3 рассмотрен специальный класс задач выпуклой негладкой оптимизации, к которым применима концепция определения 3.1.1 ($\delta = \gamma = 0$, $\Delta > 0$). По сути, показан смысл параметров неточности предложенных подходов ко введению оптимизационной модели для некоторого класса негладких задач. Доказано, что для таких задач возможно модифицировать алгоритмы 6 и 7 так, чтобы гарантированно имела место сходимость по функции со скоростью $O\left(\varepsilon^{-2}\log_2\varepsilon^{-1}\right)$, которая близка к оптимальной на классе задач выпуклой негладкой оптимизации. Экспериментально для некоторых примеров негладких задач показано, что за счёт адаптивности алгоритмов 6 и 7 может наблюдаться существенно более высокая скорость сходимости $O(\varepsilon^{-1})$ или даже выше, что свойственно выпуклым гладким задачам.

– В разделе 3.4 рассмотрены варианты концепции (δ, Δ, L)-модели для вариационных неравенств и седловых задач, получены оценки скорости сходимости адаптивных вариантов проксимального зеркального метода, также проведены вычислительные эксперименты, иллюстрирующие преимущества адаптивных подходов.

3.1 Понятие $(\delta, \gamma, \Delta, L)$-модели функции в запрошенной точке и оценка скорости сходимости адаптивного градиентного метода для задач, допускающих существование такой модели

Введем анонсированный выше аналог понятия $(\delta, \gamma, \Delta, L)$-модели целевой функции, который учитывает погрешность Δ задания градиента и

применим также для задач с относительно гладкими целевыми функционалами [124].

Определение 3.1.1. Будем говорить, что f допускает $(\delta, \gamma, \Delta, L)$-модель в точке $x \in Q$, если для некоторой выпуклой по первой переменной функции $\psi(y, x)$ такой, что $\psi(x, x) = 0$ для произвольного $y \in Q$, будет верно неравенство

$$\begin{aligned} f_\delta(x) + \psi(y, x) - \gamma \|y - x\| \leqslant \\ \leqslant f(y) \leqslant f_\delta(x) + \psi(y, x) + \delta + \Delta \|y - x\| + LV(y, x) \end{aligned} \tag{3.5}$$

для некоторого $f_\delta : Q \to \mathbb{R}$.

Покажем пример, поясняющий смысл использования модельной общности в предыдущем определении.

Пример 3.1.2. Отметим задачу (раздел 1.4.1) выпуклой композитной оптимизации $f(x) = g(x) + h(x) \to \min$, где g — гладкая выпуклая функция, а h — выпуклая не обязательно гладкая функция простой структуры (операция проектирования на любое множество уровня h не сильно затратна). Если при этом для градиента ∇g задано его приближение $\widetilde{\nabla} g$: $\left\| \widetilde{\nabla} g(x) - \nabla g(x) \right\|_* \leqslant \Delta$ и

$$g(y) \geqslant g(x) + \langle \widetilde{\nabla} g(x), y - x \rangle - \Delta \|y - x\| - \delta,$$

то можно положить

$$\psi(y, x) = \langle \widetilde{\nabla} g(x), y - x \rangle + h(y) - h(x)$$

и будет верно (3.5) при $\gamma = \Delta$.

Напомним вспомогательное утверждение (см., например, лемму 1 из [22]).

Лемма 3.1.3. *Пусть $\psi(x)$ выпуклая функция и*

$$y = \arg\min_{x \in Q} \{\psi(x) + V(x, z)\}.$$

Тогда выполнено неравенство

$$\psi(x) + V(x, z) \geqslant \psi(y) + V(y, z) + V(x, y) \quad \forall x \in Q.$$

Алгоритм 6 Адаптивный градиентный метод для функций, допускающих $(\delta, \gamma, \Delta, L)$-модель в запрошенной точке.

Require: x^0 — начальная точка, $V(x_*, x^0) \leqslant R^2$, параметры $\delta_0,\ L_0,\ \Delta_0$ $(\delta_0 \leqslant 2\delta,\ L_0 \leqslant 2L,\ \Delta_0 \leqslant 2\Delta)$.

1: $L_{k+1} := {}^{L_k}/_2$, $\Delta_{k+1} := {}^{\Delta_k}/_2$, $\delta_{k+1} := {}^{\delta_k}/_2$.

2: $x^{k+1} := \arg\min\limits_{x \in Q}\{\psi(x, x^k) + LV(x, x^k)\}$.

3: **if** $f_\delta(x^{k+1}) \leqslant f_\delta(x^k) + \psi(x^{k+1}, x^k) + L_{k+1}V(x^{k+1}, x^k) + \Delta_{k+1}\left\|x^{k+1} - x^k\right\| + \delta_{k+1}$ **then**

4: $k := k + 1$ и выполнение п. 1.

5: **else**

6: $L_{k+1} := 2 \cdot L_{k+1}$; $\Delta_{k+1} := 2 \cdot \Delta_{k+1}$; $\delta_{k+1} := 2 \cdot \delta_{k+1}$ и выполнение п. 2.

7: **end if**

Ensure: $\widehat{x} := \frac{1}{S_N}\sum\limits_{k=0}^{N-1}\frac{x^{k+1}}{L_{k+1}}$, $S_N := \sum\limits_{k=0}^{N-1}\frac{1}{L_{k+1}}$.

Теорема 3.1.4. *Пусть $f : Q \to \mathbb{R}$ — выпуклая функция, $\psi(y, x)$ — $(\delta, \gamma, \Delta, L)$-модель на Q и $V(x_*, x^0) \leqslant R^2$, где x^0 — начальное приближение, а x_* — точное решение, ближайшее к x^0 с точки зрения дивергенции Брегмана. Тогда после N итераций для выхода $\widehat{x}$ алгоритма 6 будет верно неравенство*

$$f(\widehat{x}) - f(x_*) \leqslant \frac{R^2}{S_N} + \frac{1}{S_N}\sum_{k=0}^{N-1}\frac{\delta_k + \Delta_k\left\|x^{k+1} - x^k\right\| + \gamma\left\|x^k - x_*\right\|}{L_{k+1}} + \delta. \tag{3.6}$$

Отметим, что вспомогательная задача п. 2 листинга алгоритма 6 решается не более

$$2N + \max\left\{\log_2\frac{2L}{L_0}, \log_2\frac{2\delta}{\delta_0}, \log_2\frac{2\Delta}{\Delta_0}\right\} \tag{3.7}$$

раз.

Доказательство. 1) Согласно лемме 3.1.3 после завершения k-й итерации $(k = 0, 1, 2, ...)$ алгоритма 6 будут верны неравенства:

$$\psi(x^{k+1}, x^k) \leqslant \psi(x, x^k) + L_{k+1}V(x, x^k) - L_{k+1}V(x, x^{k+1}) - L_{k+1}V(x^{k+1}, x^k),$$

$$f_\delta(x^{k+1}) \leqslant f_\delta(x^k) + \psi(x^{k+1}, x^k) + L_{k+1}V(x, x^k) - L_{k+1}V(x, x^{k+1}) +$$

$$+\Delta_{k+1}\left\|x^{k+1}-x^k\right\|+\delta_{k+1}.$$

Поэтому

$$f_\delta(x^{k+1})\leqslant f_\delta(x^k)+\psi(x,x^k)+L_{k+1}V(x,x^k)-L_{k+1}V(x,x^{k+1})+$$

$$+\Delta_{k+1}\left\|x^{k+1}-x^k\right\|+\delta_{k+1}.$$

Далее, с учетом левой части неравенства (3.5) при $x=x_*$ получим

$$f(x^{k+1})-f(x_*)\leqslant L_{k+1}V(x_*,x^k)-L_{k+1}V(x_*,x^{k+1})+$$

$$+\Delta_{k+1}\left\|x^{k+1}-x^k\right\|+\delta_{k+1}+\delta+\gamma\left\|x^k-x_*\right\|,$$

откуда после суммирования по $k=0,1,...,N-1$ ввиду выпуклости f имеем:

$$f(\widehat{x})-f(x_*)\leqslant\frac{1}{S_N}\sum_{k=0}^{N-1}\frac{f(x^{k+1})}{L_{k+1}}-f(x_*)\leqslant V(x_*,x^0)+$$

$$+\frac{1}{S_N}\sum_{k=0}^{N-1}L_{k+1}^{-1}(\Delta_{k+1}\left\|x^{k+1}-x^k\right\|+\delta_{k+1}+\gamma\left\|x^k-x_*\right\|)+\delta.$$

2) Проверим оценку (3.7). Пусть на $(k+1)$-й итерации ($k=0,1,...,N-1$) алгоритма 6 вспомогательная задача решается i_{k+1} раз. Тогда

$$2^{i_{k+1}-2}=\frac{L_{k+1}}{L_k}=\frac{\delta_{k+1}}{\delta_k}=\frac{\Delta_{k+1}}{\Delta_k},$$

поскольку в начале каждой итерации параметры L_k,δ_k,Δ_k делятся на 2. Поэтому

$$\sum_{k=0}^{N-1}i_{k+1}=2N+\log_2\frac{L_N}{L_0},\quad\log_2\frac{L_N}{L_0}=\log_2\frac{\delta_N}{\delta_0}=\log_2\frac{\Delta_N}{\Delta_0}.$$

Ясно, что верно хотя бы 1 из неравенств $L_N\leqslant 2L$, $\delta_N\leqslant 2\delta$ и $\Delta_N\leqslant 2\Delta$, что и обосновывает (3.7). □

Замечание 3.1.5. Оценка (3.7) показывает, что в среднем трудоемкость итерации предложенного адаптивного алгоритма превышает трудоемкость неадаптивного метода не более, чем в постоянное число раз. Отметим также, что при $k=0,1,2,...$ $L_{k+1}\leqslant 2CL$, $C=$

$\max\left\{1, \frac{2\delta}{\delta_0}, \frac{2\Delta}{\Delta_0}\right\}$. Поэтому $S_N \leqslant \frac{N}{2CL}$, что указывает на скорость сходимости метода $O(\varepsilon^{-1})$, но при наличии в оценке (3.6) слагаемых, определяемой параметрами δ, γ, Δ (при этом ввиду адаптивности метода δ_k и Δ_k могут быть меньше δ и Δ соотвественно). Можно доказать, что эта величина ограничена в случае ограниченного допустимого множества задачи Q, что вполне может считаться приемлемым [46].

Замечание 3.1.6. Отметим, что ввиду адаптивности алгоритма 6, полученная в теореме 3.1.4 оценка скорости сходимости может быть применена даже в случаях $L = +\infty$, $\Delta = +\infty$ или $\delta = +\infty$. Если не происходит зацикливание и каждый раз выполняется критерий выхода из итерации, то алгоритм 6 применим и в этом случае.

3.2 Оценка скорости сходимости для ускоренного градиентного метода для задач минимизации функционалов, допускающих существование (δ, Δ, L)-модели целевой функции в произвольной запрошенной точке

Данный раздел посвящён обоснованию теоретической оценки скорости сходимости адаптивного варианта быстрого градиентного метода Ю. Е. Нестерова (метод подобных треугольников) для класса задач, аналогичного выделенному в предыдущем пункте классе задач. Смысл использования этого метода в том, что при отсутствии погрешностей он приводит к оптимальным оценкам скорости сходимости на классе задач выпуклой оптимизации с липшицевым градиентом. Однако для ускоренного варианта необходимо несколько сузить класс целевых функционалов по сравнению с определением 3.2.1. В частности, здесь уже важна 1-сильная выпуклость прокс-функции и дивергенции Брегмана.

Определение 3.2.1. Будем говорить, что f допускает (δ, Δ, L)-модель в точке $x \in Q$, если для некоторой выпуклой по первой переменной функции $\psi(y, x)$ такой, что $\psi(x, x) = 0$ для произвольного $y \in Q$, будет

Алгоритм 7 Ускоренный градиентный метод с оракулом, использующим (δ, Δ, L)-модель в запрошенной точке.

Require: $x^0 \in Q$ — начальная точка, $V(x_*, x^0) \leqslant R^2$, параметры $L_0 > 0,\ \Delta_0 > 0,\ \delta_0 > 0$
$(L_0 \leqslant 2L,\ \Delta_0 \leqslant 2\Delta,\ \delta_0 \leqslant 2\delta)$.

0-шаг: $y^0 := x^0$, $u^0 := x^0$, $L_1 := \frac{L_0}{2}$, $\Delta_1 := \frac{\Delta_0}{2}$, $\delta_1 := \frac{\delta_0}{2}$, $\alpha_0 := 0$, $A_0 := \alpha_0$

for $k = 1, \ldots$ **do**

Находим наибольший корень α_{k+1}:

$$A_k + \alpha_{k+1} = L_{k+1}\alpha_{k+1}^2;$$

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

$$\varphi_{k+1}(x) = V(x, u^k) + \alpha_{k+1}\psi(x, y^{k+1});$$

$$u^{k+1} := \arg\min_{x \in Q} \varphi_{k+1}(x),\ x^{k+1} := \frac{\alpha_{k+1}u^{k+1} + A_k x^k}{A_{k+1}};$$

if $f_\delta(x^{k+1}) \leqslant f_\delta(y^{k+1}) + \psi(x^{k+1}, y^{k+1}) + \frac{L_{k+1}}{2}\|x^{k+1} - y^{k+1}\|^2 + \Delta_{k+1}\|x^{k+1} - y^{k+1}\| + \delta_{k+1}$ **then**

$L_{k+2} := \frac{L_{k+1}}{2}$, $\Delta_{k+2} := \frac{\Delta_{k+1}}{2}$ и $\delta_{k+2} := \frac{\delta_{k+1}}{2}$ и перейти к следующему шагу

else

$L_{k+1} := 2L_{k+1}$, $\Delta_{k+1} := 2\Delta_{k+1}$ и $\delta_{k+1} := 2\delta_{k+1}$ и повторить текущий шаг

end if

end for

верно неравенство

$$f_\delta(x) + \psi(y,x) \leqslant f(y) \leqslant f_\delta(x) + \psi(y,x) + \delta + \Delta\|y-x\| + \frac{L}{2}\|y-x\|^2 \leqslant$$

$$\leqslant f_\delta(x) + \psi(y,x) + \delta + \Delta\|y-x\| + LV(y,x)$$

для некоторого $f_\delta : Q \to \mathbb{R}$.

Замечание 3.2.2. Аналогично замечанию 3.1.5, для всякого $k \geqslant 0$ при некотором $C \geqslant 1$ выполнено $L_k \leqslant 2CL$. Для $k = 0$ верно из того, что $L_0 \leqslant 2L$. Для $k \geqslant 1$ это следует из того, что мы выйдем из внутреннего цикла, где подбирается L_k, ранее, чем L_k станет больше CL. Выход из цикла гарантируется тем, что по условию существует (δ, Δ, L)-модель для $f(x)$ в любой точке $x \in Q$.

Мы отправляемся от [22], где предложен адаптивный быстрый градиентный метод с оракулом, использующим (δ, L)-модель целевой функции в запрошенной точке. Получим оценку скорости сходимости алгоритма 7. Докажем вспомогательное утверждение.

Лемма 3.2.3. *Для произвольной точки $x \in Q$ выполнено неравенство:*

$$A_{k+1}f(x^{k+1}) - A_k f(x^k) + V(x,u^{k+1}) - V(x,u^k) \leqslant$$

$$\leqslant \alpha_{k+1}f(x) + (\Delta_{k+1}\left\|x^{k+1} - y^{k+1}\right\| + \delta_{k+1} + \delta)A_{k+1}.$$

Доказательство. Рассмотрим цепочку неравенств:

$$f(x^{k+1}) \leqslant f_\delta(y^{k+1}) + \psi(x^{k+1}, y^{k+1}) + \frac{L_{k+1}}{2}\left\|x^{k+1} - y^{k+1}\right\|^2 +$$

$$+\Delta_{k+1}\left\|x^{k+1} - y^{k+1}\right\| + \delta_{k+1} + \delta =$$

$$= f_\delta(y^{k+1}) + \psi\left(\frac{\alpha_{k+1}u^{k+1} + A_k x^k}{A_{k+1}}, y^{k+1}\right) +$$

$$+\frac{L_{k+1}}{2}\left\|\frac{\alpha_{k+1}u^{k+1} + A_k x^k}{A_{k+1}} - y^{k+1}\right\|^2 + \Delta_{k+1}\left\|x^{k+1} - y^{k+1}\right\| + \delta_{k+1} + \delta \leqslant$$

$$\leqslant f_\delta(y^{k+1}) + \frac{\alpha_{k+1}}{A_{k+1}}\psi(u^{k+1}, y^{k+1}) +$$

$$+\frac{A_k}{A_{k+1}}\psi(x_k, y_{k+1}) + \frac{L_{k+1}\alpha_{k+1}^2}{2A_{k+1}^2}\left\|u^{k+1} - u^k\right\|^2 +$$

$$+\Delta_{k+1}\left\|x^{k+1}-y^{k+1}\right\|+\delta_{k+1}+\delta=$$

$$=\frac{A_k}{A_{k+1}}(f_\delta(y^{k+1})+\psi(x_k,y^{k+1}))+\frac{\alpha_{k+1}}{A_{k+1}}(f_\delta(y^{k+1})+\psi(u^{k+1},y^{k+1}))+$$

$$+\frac{L_{k+1}\alpha_{k+1}^2}{2A_{k+1}^2}\left\|u^{k+1}-u^k\right\|^2+\Delta_{k+1}\left\|x^{k+1}-y^{k+1}\right\|+\delta_{k+1}+\delta=\textcircled{1}=$$

$$=\frac{A_k}{A_{k+1}}(f_\delta(y^{k+1})+\psi(x^k,y^{k+1}))+$$

$$+\frac{\alpha_{k+1}}{A_{k+1}}(f_\delta(y^{k+1})+\psi(u^{k+1},y^{k+1})+\frac{1}{2\alpha_{k+1}}\left\|u^{k+1}-u^k\right\|^2)+$$

$$+\Delta_{k+1}\left\|x^{k+1}-y^{k+1}\right\|+\delta_{k+1}+\delta\leqslant$$

$$\leqslant\frac{A_k}{A_{k+1}}(f_\delta(y^{k+1})+\psi(x^k,y^{k+1}))+$$

$$+\frac{\alpha_{k+1}}{A_{k+1}}(f_\delta(y^{k+1})+\psi(u^{k+1},y^{k+1})+\frac{1}{\alpha_{k+1}}V(u^{k+1},u^k))+$$

$$+\Delta_{k+1}\left\|x^{k+1}-y^{k+1}\right\|+\delta_{k+1}+\delta\leqslant\textcircled{2}\leqslant$$

$$\leqslant\frac{A_k}{A_{k+1}}f(x^k)+\frac{\alpha_{k+1}}{A_{k+1}}(f_\delta(y^{k+1})+\psi(x,y^{k+1})+$$

$$+\frac{1}{\alpha_{k+1}}V(x,u^k)-\frac{1}{\alpha_{k+1}}V(x,u^{k+1}))+$$

$$+\Delta_{k+1}\left\|x^{k+1}-y^{k+1}\right\|+\delta_{k+1}+\delta\leqslant$$

$$\leqslant\frac{A_k}{A_{k+1}}f(x^k)+\frac{\alpha_{k+1}}{A_{k+1}}f(x)+$$

$$+\frac{1}{A_{k+1}}V(x,u^k)-\frac{1}{A_{k+1}}V(x,u^{k+1})+\Delta_{k+1}\left\|x^{k+1}-y^{k+1}\right\|+\delta_{k+1}+\delta.$$

Неравенство ① следует из равенства $A_k=L_k\alpha_k^2$. Неравенство ② следует из леммы 3.1.3 с $\psi(x)=\alpha_{k+1}\psi(x,y^{k+1})$ и $f(x)\geqslant f_\delta(y)+\psi(x,y)$. □

Теорема 3.2.4. *Пусть $V(x_*,x^0)\leqslant R^2$, где x^0 — начальная точка, а x_* — ближайшая точка минимума к точке x^0 в смысле дивергенции Брегмана. Для предложенного алгоритма 7 после N итераций выполнено следующее неравенство:*

$$f(x^N)-f(x_*)\leqslant\frac{R^2}{A_N}+\frac{1}{A_N}\sum_{k=0}^{N-1}(\Delta_{k+1}\left\|x^{k+1}-y^{k+1}\right\|+\delta_{k+1}+\delta)A_{k+1}.$$

Отметим, что вспомогательная задача пункта 6 листинга алгоритма 7 решается не более

$$2N + \max\left\{\log_2 \frac{2L}{L_0}, \log_2 \frac{2\delta}{\delta_0}, \log_2 \frac{2\Delta}{\Delta_0}\right\} \tag{3.8}$$

раз.

Доказательство. Просуммировав неравенства из леммы 3.2.3 по $k = 0, ..., N-1$, получим

$$A_N f(x^N) - A_0 f(x^0) + V(x, u^N) - V(x, u^0) \leqslant (A_N - A_0) f(x) +$$

$$+ \sum_{k=0}^{N-1} A_{k+1}(\Delta_{k+1} \left\|x^{k+1} - y^{k+1}\right\| + \delta_{k+1} + \delta),$$

откуда

$$A_N f(x^N) - A_0 f(x^0) + V(x, u^N) - V(x, u^0) \leqslant (A_N - A_0) f(x) +$$

$$+ \sum_{k=0}^{N-1} A_{k+1}(\Delta_{k+1} \left\|x^{k+1} - y^{k+1}\right\| + \delta_{k+1} + \delta),$$

Если выбрать $x = x_*$, то

$$A_N(f(x^N) - f(x_*) \leqslant R^2 + 2\sum_{k=0}^{N-1} A_{k+1}(\Delta_{k+1} \left\|x^{k+1} - y^{k+1}\right\| + \delta_{k+1} + \delta).$$

Разделим обе части неравенства на A_N и тогда получим, что

$$f(x^N) - f(x_*) \leqslant \frac{R^2}{A_N} + \frac{2\sum_{k=0}^{N-1}(\Delta_{k+1} \left\|x^{k+1} - y^{k+1}\right\| + \delta_{k+1} + \delta)A_{k+1}}{A_N}$$

Оценка (3.8) проверяется аналогично пункту 2 доказательства теоремы 3.1.4. □

Полученную оценку можно несколько конкретизировать с использованием следующего вспомогательного утверждения.

Лемма 3.2.5. *Пусть для последовательности α_k верно*

$$\alpha_0 = 0, \;\; A_k = \sum_{i=0}^{k} \alpha_i, \;\; A_k = L_k \alpha_k^2,$$

где для фиксированного $C \geqslant 1$ верно $L_k \leqslant 2CL$ при всяком $k \geq 0$ согласно замечанию 3.2.2 выше. Тогда для любого $k \geqslant 1$ верно следующее неравенство

$$A_k \geqslant \frac{(k+1)^2}{8CL}.$$

Доказательство. Проведём доказательство индукцией по k аналогично [22]. Пусть $k = 1$. Тогда $\alpha_1 = L_1\alpha_1^2$ и

$$A_1 = \alpha_1 = \frac{1}{L_1} \geqslant \frac{1}{2CL}.$$

Пусть $k \geqslant 2$. В таком случае верны следующие эквивалентные равенства:

$$L_{k+1}\alpha_{k+1}^2 = A_{k+1} \Leftrightarrow L_{k+1}\alpha_{k+1}^2 = A_k + \alpha_{k+1} \Leftrightarrow L_{k+1}\alpha_{k+1}^2 - \alpha_{k+1} - A_k = 0.$$

Выберем больший корень указанного квадратного уравнения:

$$\alpha_{k+1} = \frac{1 + \sqrt{1 + 4L_{k+1}A_k}}{2L_{k+1}}.$$

По индукции допустим, что неравенство (3.2.5) верно для k, тогда:

$$\alpha_{k+1} = \frac{1}{2L_{k+1}} + \sqrt{\frac{1}{4L_{k+1}^2} + \frac{A_k}{L_{k+1}}} \geqslant \frac{1}{2L_{k+1}} + \sqrt{\frac{A_k}{L_{k+1}}} \geqslant$$
$$\geqslant \frac{1}{4CL} + \frac{1}{\sqrt{2CL}} \frac{k+1}{4\sqrt{CL}} = \frac{k+2}{4CL}.$$

Поэтому

$$\alpha_{k+1} \geqslant \frac{k+2}{4CL}$$

и

$$A_{k+1} = A_k + \alpha_{k+1} = \frac{(k+1)^2}{8CL} + \frac{k+2}{4CL} \geqslant \frac{(k+2)^2}{8CL}.$$

□

Таким образом, из теоремы 3.2.4 вытекает

Следствие 3.2.6. *Пусть $V(x_*, x^0) \leqslant R^2$, где x^0 — начальная точка, а x_* — ближайшая точка минимума к точке x^0 в смысле дивергенции*

Брегмана. Для предложенного алгоритма выполнено следующее неравенство:

$$f(x^N)-f(x_*) \leqslant \frac{4CLR^2}{(N+1)^2}+\frac{1}{A_N}\sum_{k=0}^{N-1}(\Delta_{k+1}\left\|x^{k+1}-y^{k+1}\right\|+\delta_{k+1}+\delta)A_{k+1}.$$

Рассмотрим пример расчётов для задачи нахождения приближённого решения матричного уравнения для предложенного выше варианта быстрого градиентного метода (алгоритм 7) с адаптацией величин, соответствующих погрешностям.

Пример 3.2.7. Пусть имеется матрица A размера 1000×1000 и вектор $b\in\mathbb{R}^{1000}$. Предположим, что главная диагональ A содержит случайно подобранные целые числа из интервала $[1,1000]$. Также другие 100 случайно выбранных элементов этой матрицы заменены целыми числами из интервала $[1,1000]$. Рассмотрим задачу решения матричного уравнения $Ax=b$. Если эта задача разрешима, то её можно сводить к задаче минимизации выпуклого функционала

$$f(x)=\frac{1}{2}\|Ax-b\|_2^2-\widehat{\delta},$$

где $0\leqslant\widehat{\delta}\leqslant\delta$. Выбранная целевая функция (при $\widehat{\delta}$) имеет L-липшицев градиент (L — наибольшее собственное значение A^TA, где A^T — матрица, транспонированная в A). Начальное приближение выберем в точке $x^0=(0,\ldots,0)$ и будем задавать случайно погрешности задания целевого функционала и градиента во всякой текущей точке, ограничив их по норме величиной $\delta=\Delta=\frac{1}{20}$. В качестве Q выберем единичный шар и сравним работу методов с адаптивной настройкой δ и Δ с работой без нее:

$$f(x^N)-f(x_*)\leqslant\frac{R^2}{A_N}+\frac{1}{A_N}\sum_{k=0}^{N-1}(\Delta_{k+1}\left\|x^{k+1}-y^{k+1}\right\|+\delta_{k+1}+\delta)A_{k+1}$$

и

$$f(x^N)-f(x_*)\leqslant\frac{R^2}{A_N}+\frac{1}{A_N}\sum_{k=0}^{N-1}(\Delta\left\|x^{k+1}-y^{k+1}\right\|+2\delta)A_{k+1},$$

которые определяют качество решения для рассмотренного выше варианта быстрого градиентного метода и его неадаптивной версии без

Таблица 3.1. Адаптивный вариант ускоренного градиентного метода для задачи решения матричного уравнения.

K	Адаптивный		Неадаптивный	
	Оценка	Время, с	Оценка	Время, с
10	1517268.8244	0.5	26187.69085	0.2
20	257427.56825	1	376.83433	0.5
30	28527.10388	1.5	52.69375	1
40	3454.98299	2	7.8553	1.6
50	587.87202	2.5	1.64234	2
100	0.81657	5	0.54552	4
200	0.27186	9	0.5397	9
300	0.28834	14	0.54792	14
400	0.42155	19	0.54195	19
500	0.24924	24	0.54839	24
600	0.24819	28	0.54359	29
700	0.78957	33	0.53985	33
800	0.80511	38	0.55039	38
900	0.3628	43	0.54324	43
1000	0.28103	48	0.54505	48

адаптивной настройки на величины погрешностей. Усредненные результаты пяти экспериментов представлены в таблице 3.1. В проведенных экспериментах значения целевой функции принадлежат промежутку $\left(85.95 \cdot 10^6, 86.08 \cdot 10^6\right)$. Поскольку значения целевого функционала получаются большими, оценивать качество решения возможно лишь согласно теоретическим оценкам, которые лучше для адаптивного варианта метода. Полученные результаты означают, что при таких данных уравнение $Ax = b$ неразрешимо на выбранном допустимом множестве.

Как видим, адаптивная настройка на величины погрешностей может позволить примерно в 2 раза улучшить гарантированное качество решения с точки зрения теоретических оценок.

3.3 О скорости сходимости методов с адаптацией к величинам погрешностей для одного класса негладких задач

В этом пункте мы рассмотрим один из основных результатов настоящей главы. Покажем, как возможно модифицировать разработанные в предыдущих двух пунктах методы, чтобы в оценках скорости сходимости не было накопления величин, соответствующих параметру Δ введенных нами выше понятий неточной оптимизационной модели. Оказывается, возможны различные интерпретации параметра Δ. Первым вариантом интерпретации может служить погрешность задания градиента для достаточно гладких задач, во втором варианте параметр Δ можно понимать как характеристику отклонения от гладкости функционала f. Точнее говоря, Δ можно понимать, например, как верхнюю оценку суммы диаметров субдифференциалов f в точках негладкости вдоль всевозможных векторных отрезков $[x;y]$ из области определения f. Рассмотрим следующий класс негладких выпуклых функционалов ([52, 55]).

Определение 3.3.1. Будем говорить, что выпуклый функционал $f\colon Q \to \mathbb{R}$ ($Q \subset \mathbb{R}^n$) имеет (δ, L)-липшицев субградиент ($f \in C^{1,1}_{L,\Delta}(Q)$), если:

(i) для произвольных $x, y \in Q$ f дифференцируем во всех точках множества $\{y_t\}_{0 \leqslant t \leqslant 1}$, за исключением последовательности (возможно, конечной)

$$\{y_{t_j}\}_{j=1}^{\infty} : \ t_1 < t_2 < t_3 < \ldots \text{ и } \lim_{j \to \infty} t_j = 1; \tag{3.9}$$

(ii) для последовательности точек из (3.9) существуют конечные субдифференциалы в смысле выпуклого анализа $\{\partial f(y_{t_j})\}_{j=1}^{\infty}$ и

$$diam\, \partial f(y_{t_j}) =: \Delta_j > 0, \text{ где } \sum_{j=1}^{+\infty} \Delta_j =: \Delta < +\infty. \tag{3.10}$$

$$(diam\, \partial f(x) = \max\{\|y - z\|_* \mid y, z \in \partial f(x)\});$$

(iii) для произвольных $x, y \in Q$ при условии, что $y_t \in Q \backslash Q_0$ при всяком $t \in (0,1)$ (то есть существует градиент $\nabla f(y_k)$) для некоторой

фиксированной константы $L > 0$, не зависящей от выбора x и y, выполняется неравенство:

$$\min_{\nabla f(x)\in\partial f(x),\ \nabla f(y)\in\partial f(y)} \|\nabla f(x) - \nabla f(y)\|_* \leqslant L\|x - y\|.$$

Ясно, что всякая выпуклая функция, удовлетворяющая (5.4), будет входить в класс $C^{1,1}_{L,\Delta}(Q)$ при $\Delta = 0$. Приведем пример негладкой функции $f \in C^{1,1}_{L,\Delta}(Q)$ при $\Delta > 0$.

Пример 3.3.2. Зафиксируем некоторое $k > 0$, величину $\Delta > 0$ и рассмотрим кусочно-линейную функцию $f : [0;1] \to \mathbb{R}$ (здесь $Q = [0;1] \subset \mathbb{R}$):

$$f(x) := kx \text{ при } 0 \leqslant x \leqslant \frac{1}{2}, \tag{3.11}$$

$$f(x) := \left(k + \sum_{i=1}^{n} \frac{\Delta}{2^i}\right) x - \sum_{i=1}^{n} \frac{\Delta}{2^i}\left(1 - \frac{1}{2^i}\right) \text{ при } 1 - \frac{1}{2^n} < x \leqslant 1 - \frac{1}{2^{n+1}},$$

$$f(1) := \lim_{x\to+1} f(x).$$

В этом случае

$$Q_0 = \left\{1 - \frac{1}{2^n}\right\}_{n=1}^{\infty},$$

$$\partial f(q_n) = \left[k + \sum_{i=1}^{n-1} \frac{\Delta}{2^i};\ k + \sum_{i=1}^{n} \frac{\Delta}{2^i}\right]$$

при $n > 1$,

$$\partial f(q_1) = \left[k; k + \frac{\Delta}{2}\right]$$

(здесь $q_n = 1 - \frac{1}{2^n}$ при $n = 1, 2, 3, \ldots$). Ясно, что диаметр $\partial f(q_n)$ равен $\frac{\Delta}{2^n}$, то есть верно (3.10) для введенной величины $\Delta > 0$. При этом на отрезках $(q_n; q_{n+1})$, $(0; q_1)$ f имеет липшицев градиент с константой $L = 0$. Поэтому для функции f из (3.11) $f \in C^{1,1}_{0,\Delta}(Q)$.

Замечание 3.3.3. Ясно, что функцию f из (3.11) нельзя представить в виде максимума конечного набора линейных функций ввиду бесконечного набора точек недифференцируемости f.

Сформулируем для введенного класса $C^{1,1}_{L,\Delta}(Q)$ следующий аналог стандартного утверждения о квадратичной интерполяции для гладких функционалов.

Теорема 3.3.4. *Пусть $f \in C^{1,1}_{L,\Delta}(Q)$. Тогда для произвольных $x, y \in Q$ верно неравенство*

$$|f(y) - f(x) - \langle \widehat{\partial} f(x), y - x\rangle| \leqslant \frac{L}{2}\|y - x\|^2 + \Delta\|y - x\|$$

для некоторого субградиента $\widehat{\partial} f(x) \in \partial f(x)$.

Доказательство. Для $x, y \in Q$ через y_t будем обозначать элемент $ty + (1-t)x$. Тогда можно ввести выпуклую функцию $\varphi : [0;1] \to \mathbb{R}$: $\varphi(t) := f(y_t)$ ($\varphi(0) = f(x)$, а $\varphi(1) = f(y)$). Поскольку для всякой точки y_t ($t \in [0;1]$) существует конечный субдифференциал $\partial f(y_t)$, то для всех $t \in (0;1)$ существуют конечные левосторонняя и правосторонняя производные:

$$\varphi'_-(t) = \lim_{\Delta t \to -0} \frac{\varphi(t + \Delta t) - \varphi(t)}{\Delta t},$$
$$\varphi'_+(t) = \lim_{\Delta t \to +0} \frac{\varphi(t + \Delta t) - \varphi(t)}{\Delta t},$$
причем $\varphi'_-(t) \leqslant \varphi'_+(t)$ и

$$\varphi'_+(t) = \max_{\widehat{\partial} f(y_t) \in \partial f(y_t)} \langle \widehat{\partial} f(y_t), y - x\rangle \text{ —} \tag{3.12}$$

производная f по направлению $y - x$ в точке y_t. Ясно, что при $y_t \notin Q_0$ (то есть существует градиент $\nabla f(y_t)$)

$$\varphi'_-(t) = \varphi'_+(t) = \langle \nabla f(y_t), y - x\rangle.$$

Ввиду выпуклости f (и φ) функция $\varphi'_+(t)$ возрастает по t на промежутке $(0;1)$. Поэтому имеем равенства:

$$f(y) = f(x) + \int\limits_{[0;1]\setminus Q_0} \langle \nabla f(y_t), y - x\rangle\, dt = \varphi(0) + \int\limits_0^1 \varphi'_+(t)\, dt,$$

откуда для подходящего субградиента $\widehat{\partial} f(x) \in \partial f(x)$ имеем:

$$f(y) = f(x) + \langle \widehat{\partial} f(x), y - x\rangle +$$
$$+ \int\limits_0^1 \left[\max_{\widehat{\partial} f(y_t) \in \partial f(y_t)} \langle \widehat{\partial} f(y_t), y - x\rangle - \langle \widehat{\partial} f(x), y - x\rangle \right] dt = \tag{3.13}$$

$$= f(x) + \langle \widehat{\partial} f(x), y - x \rangle + \int_0^1 \langle \widehat{\partial} f(y_t) - \widehat{\partial} f(x), y - x \rangle \, dt$$

для набора субградиентов $\{\widehat{\partial} f(y_t)\}_{t \in (0;1]}$, на которых достигаются соответствующие максимумы.

Если $y_t \in Q_0$, то $y_t = q_k$ $(k \geqslant 1)$ из определения 3.3.1(ii) и тогда

$$\begin{aligned} \varphi'_+(t) - \varphi'_-(t) &= \langle \widehat{\partial}_1 f(y_t) - \widehat{\partial}_2 f(y_t), y - x \rangle = \\ &= \langle \widehat{\partial}_1 f(q_k) - \widehat{\partial}_2 f(q_k), y - x \rangle \leqslant \\ \leqslant \|\widehat{\partial}_1 f(q_k) - \widehat{\partial}_2 f(q_k)\|_* \cdot \|y - x\| &\overset{(3.10)}{\leqslant} \Delta_k \|y - x\| \end{aligned}$$

для соответствующих субградиентов $\widehat{\partial}_{1,2} f(q_k)$.

Не уменьшая общности рассуждений, будем считать, что

$$x, y \in Q_0 \subset \{y_t\}_{t \in [0;1]}$$

и всякому q_n поставим в соответствие $t_n \in [0;1] : q_n = (1 - t_n)x + t_n y$.

Пусть существует последовательность

$$\{t_n\}_{n=1}^{\infty} : 0 = t_1 < t_2 < \ldots < 1, \ \lim_{n \to \infty} t_n = 1.$$

Тогда $\forall \tau_1, \tau_2 \in (t_k; t_{k+1})$ при $k \geqslant 1$ верны неравенства:

$$\|\nabla f(y_{\tau_2}) - \nabla f(y_{\tau_1})\|_* \leqslant L|\tau_2 - \tau_1| \cdot \|y - x\|,$$

Поэтому при выборе в (3.13) подходящего субградиента $\widehat{\partial} f(x)$ можно подобрать оператор $\widehat{g} : E \to E^*$ так, что будут выполняться соотношения:

$$\left| f(y) - f(x) - \langle \widehat{\partial} f(x), y - x \rangle \right| = \left| \int_0^1 \langle \widehat{\partial} f(y_t) - \widehat{\partial} f(x), y - x \rangle \, dt \right| =$$

$$= \left| \int_0^1 (\varphi'_+(t) - \varphi'_+(0)) \, dt \right| \leqslant \left| \int_0^1 \langle \widehat{g}(y_t) - \widehat{\partial} f(x), y - x \rangle \, dt \right| + \sum_{k=1}^{+\infty} \Delta_k \|y - x\|,$$

причем

$$\|\widehat{g}(y_{\tau_1}) - \widehat{g}(y_{\tau_2})\|_* \leqslant L \|y_{\tau_1} - y_{\tau_2}\| \quad \forall \tau_1, \tau_2 \in [0;1].$$

Поэтому

$$\left|\int\limits_0^1 \langle \widehat{g}(y_t) - \widehat{\partial} f(x), y - x\rangle \, dt\right| = \left|\int\limits_0^1 \langle \widehat{g}(y_t) - \widehat{g}(x), y - x\rangle \, dt\right| \leqslant$$

$$\leqslant \int\limits_0^1 |\langle \widehat{g}(y_t) - \widehat{g}(x), y - x\rangle| \, dt \leqslant \int\limits_0^1 \|\widehat{g}(y_t) - \widehat{g}(x)\|_* \, dt \cdot \|y - x\| \leqslant$$

$$\leqslant L \int\limits_0^1 \|y_t - x\| \, dt \cdot \|y - x\| \leqslant L\|y - x\|^2 \cdot \int\limits_0^1 t \, dt = \frac{L}{2}\|y - x\|^2,$$

то есть

$$|f(y) - f(x) - \langle \widehat{\partial} f(x), y - x\rangle| \leqslant \frac{L}{2}\|y - x\|^2 + \Delta\|y - x\|,$$

что и требовалось доказать. □

Поскольку субградиенты f в одной и той же точке отличаются не более, чем на Δ, то верно

Следствие 3.3.5. *Всякий функционал $f \in C^{1,1}_{L,\Delta}(Q)$ для произвольного субградиента $\nabla f(x) \in \partial f(x)$ удовлетворяет неравенству*

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 + 2\Delta\|y - x\| \quad \forall y \in Q.$$

С другой стороны, ввиду выпуклости f будет верно $f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle$. Поэтому для всякой функции $f \in C^{1,1}_{L,\Delta}(Q)$ существует $(0, 2\Delta, 0, L)$-модель согласно определению 3.1.1. Точнее говоря, $\psi(y, x) = \langle \nabla f(x), y - x\rangle$.

Оказывается, что за счет адаптивности экспериментально скорость сходимости метода может оказаться выше, чем по отмеченной выше оценке. Приведем некоторые примеры. Здесь для оценки качества решения будем использовать теоретические оценки, поскольку постановка задачи не указывает явно на оценку искомого минимального значения.

Задача 3.3.6. (Задача Ферма–Торричелли–Штейнера) Пусть в пространстве $\mathbb{R}^n$ (размерность $n = 10^5$) задано несколько точек и нужно найти такую, для которой сумма евклидовых расстояний до данных точек минимальна. Для заданных 10 точек $A_k = (a_{1k}, a_{2k}, \ldots, a_{nk})$

Таблица 3.2. Результаты численных экспериментов (задача 3.3.6).

Итерации	200	400	600	800	1000
Оценка	0.03715	0.01863	0.01250	0.00945	0.00762
Время, с	29	58	87	115	139

$(k = 1, 2, \ldots, 10)$ (координаты точек A_k представляют выбираются случайно так, чтобы $0.5 < \sqrt{a_{1k}^2 + a_{2k}^2 + \ldots + a_{nk}^2} < 1,\ k = \overline{1,m},\ m = 10$) в n-мерном евклидовом пространстве $\mathbb{R}^n$ необходимо найти такую точку $X = (x_1, x_2, \ldots, x_n)$, чтобы целевая функция

$$f(x) := \sum_{k=1}^{m} XA_k = \sum_{k=1}^{m} \sqrt{(x_1 - a_{1k})^2 + (x_2 - a_{2k})^2 + \ldots + (x_n - a_{nk})^2}$$

принимала наименьшее значение на единичном шаре с центром в нуле. В таблице 3.2 приведены усредненные результаты 10 экспериментов со случайным выбором координат точек для указанного количества итераций.

Как видим, скорость сходимости метода (по теоретической оценке) получается близка к $O(\varepsilon^{-1})$. Это свойственно для градиентных методов на классе задач оптимизации выпуклых функций с липшицевым градиентом (так называемых гладких задач). Однако рассматриваемая в данном примере задача негладкая, поскольку точки недифференцируемости f лежат в области определения (единичном шаре с центром в нуле). Для задач минимизации выпуклых липшицевых функций, как известно, оптимальная оценка скорости сходимости (суб)градиентных методов $O(\varepsilon^{-2})$ [36]. Оценку $O(\varepsilon^{-1})$ можно объяснить адаптивностью предложенного метода.

В предыдущем примере было конечное число точек негладкости. Рассмотрим теперь ещё результаты экспериментов для случая бесконечного числа точек негладкости (недифференцируемости) целевого функционала, но с конечным Δ. Отметим, что задачи указанного типа рассматривались, в частности, в [130].

Задача 3.3.7. (Аналог задачи Ферма–Торричелли–Штейнера) Для заданных шаров ω_k с центрами $A_k = (a_{1k}, a_{2k}, \ldots, a_{nk})$ (координаты точек A_k выбираются случайно так, чтобы $1 < \sqrt{a_{1k}^2 + a_{2k}^2 + \ldots + a_{nk}^2} < 1.5,\ k = \overline{1,m},\ m = 10$) и единичными радиусами в n-мерном евклидовом пространстве $\mathbb{R}^n$ $(n = 10^5)$ необходимо найти такую точку

Таблица 3.3. Результаты численных экспериментов (задача 3.3.7).

Итерации	200	400	600	800	1000
Оценка	0.02315	0.01168	0.00785	0.00595	0.00480
Время, с	27	54	82	110	136

$X = (x_1, x_2, \ldots, x_n)$, чтобы целевая функция

$$f(x) := \sum_{k=1}^{m} d(X, \omega_k)$$

принимала наименьшее значение на множестве точек единичного шара с центром в нуле, где

$$d(X, \omega_k) = \begin{cases} XA_k - 1, & \text{если } XA_k > 1; \\ 0, & \text{иначе,} \end{cases}$$

В таблице 3.3 приведены усредненные результаты 10 экспериментов со случайным подбором координат точек для указанного количества итераций. Как видим, скорость сходимости метода снова близка к $O(\varepsilon^{-1})$.

Рассмотрим ещё пример, в котором все точки некоторого векторного отрезка могут быть точками негладкости.

Задача 3.3.8. (Задача о наименьшем покрывающем шаре) Для заданных точек $A_k = (a_{1k}, a_{2k}, \ldots, a_{nk})$ найти евклидов шар наименьшего радиуса, в котором лежат эти точки. Координаты точек A_k выбираются случайно так, что $0.5 < \sqrt{a_{1k}^2 + a_{2k}^2 + \ldots + a_{nk}^2} < 1$, $k = \overline{1,10}$ в n-мерном евклидовом пространстве $\mathbb{R}^n$ (размерность $n = 10^5$) необходимо найти такую точку $X = (x_1, x_2, \ldots, x_n)$, чтобы целевая функция

$$f(x) := \max_{k=\overline{1,m}} XA_k = \max_{k=\overline{1,m}} \sqrt{(x_1 - a_{1k})^2 + (x_2 - a_{2k})^2 + \ldots + (x_n - a_{nk})^2}$$

принимала наименьшее значение на множестве точек единичного шара с центром в нуле. В таблице 3.4 приведены усредненные результаты 10 экспериментов со случайным подбором координат точек для определенного количества итераций. Как видим, скорость сходимости метода снова близка к $O(\varepsilon^{-1})$.

Таблица 3.4. Результаты численных экспериментов (задача 3.3.8).

Итерации	200	400	600	800	1000
Оценка	0.79321	0.44447	0.31116	0.24097	0.19767
Время, с	15	29	44	58	72

Приведённые результаты экспериментов указывают на неплохую эффективность предложенной адаптивной процедуры регулировки шага в методе. Однако можно в некотором смысле и теоретически показать оптимальность предложенной схемы для рассматриваемых негладких задач. Оказывается, в случае известной величины $\Delta < +\infty$ возможно несколько модифицировать алгоритм 1, обеспечив уменьшение $\Delta_{k+1}\|x^{k+1} - x^k\|$ в (3.6) до любой заданной величины. Это позволит показать оптимальность данного метода в теории нижних оракульных оценок с точностью до логарифмического множителя.

Покажем, как это возможно сделать. Пусть на (k+1)-й итерации алгоритма 6 ($k = 0, 1, \ldots, N-1$) верно неравенство $L \leqslant L_{k+1} \leqslant 2L$ (как показано в п. 2 доказательства теоремы 3.1.4, этого можно всегда добиться выполнением не более чем постоянного числа операций п. 2 листинга алгоритма 6). Для каждой итерации алгоритма 6 ($k = 0, 1, \ldots, N-1$) предложим такую процедуру:

Повторяем операции п. 2 p раз, увеличивая L_{k+1} в два раза при неизменной $\Delta_{k+1} \leqslant 2\Delta$.	(3.14)

Процедуру (3.14) остановим в случае выполнения одного из неравенств:

$$\Delta_{k+1} \left\|x^{k+1} - x^k\right\| \leqslant \frac{\varepsilon}{2}, \tag{3.15}$$

или

$$f(x^{k+1}) \leqslant f(x^k) + \left\langle \widetilde{\nabla} f(x^k), x^{k+1} - x^k \right\rangle + 2^{p-1} L \left\|x^{k+1} - x^k\right\|^2. \tag{3.16}$$

Отметим, что здесь мы полагаем f точно заданной, то есть $f_\delta = f$ ($\delta = 0$); $\widetilde{\nabla} f$ — некоторый субградиент f. Процедура (3.14) предполагает на $(k+1)$-й итерации ($k = 0, 1, 2, \ldots, N-1$) обновления x^{k+1} (при сохранении x^k). Оценим количество повторений p шага п. 2 листинга алгоритма 1, необходимое для достижения альтернативы (3.15)–(3.16). Для всяких x^k и $x^{k+1} \in Q$ по предположению верно неравенство

$$f(x^{k+1}) \leqslant f(x^k) + \left\langle \widetilde{\nabla} f(x^k), x^{k+1} - x^k \right\rangle + \frac{L}{2} \left\|x^{k+1} - x^k\right\|^2 + \Delta \left\|x^{k+1} - x^k\right\|,$$

причем $\Delta_{k+1} \leqslant 2\Delta$. Если не выполнено (3.15), то $\left\|x^{k+1} - x^k\right\| > \frac{\varepsilon}{4\Delta}$ и (3.16) заведомо верно при

$$2^p > 1 + \frac{16\Delta^2}{\varepsilon L}, \tag{3.17}$$

поскольку тогда при условии (3.15)

$$\frac{2^p - 1}{2} L \left\|x^{k+1} - x^k\right\|^2 > 2\Delta \left\|x^{k+1} - x^k\right\|.$$

Итак, после повторения p процедур (p удовлетворяет (3.17)) типа (3.14) на каждой из N итераций алгоритма 1 неравенство (3.6) примет вид:

$$f(\widehat{x}) - f^* \leqslant \frac{R^2}{S_N} + \frac{\varepsilon}{2},$$

причем

$$S_N = \sum_{k=0}^{N-1} \frac{1}{L_{k+1}} \geqslant \frac{N}{2^{p+1}L}.$$

Поэтому

$$\frac{R^2}{S_N} \leqslant \frac{2^{p+1} L R^2}{N} \leqslant \frac{\varepsilon}{2}$$

в случае $N \geqslant \frac{2^{p+1} L R^2}{2}$. С учетом (3.17) получаем оценку

$$N \geqslant \frac{4LR^2}{\varepsilon} + \frac{64\Delta^2 R^2}{\varepsilon^2}. \tag{3.18}$$

При этом (3.17) означает, что на каждой итерации потребуется $O\left(\log_2 \varepsilon^{-1}\right)$ шагов типа п. 2 листинга алгоритма 1 (то есть операций проектирования). В итоге общее количество обращений к субградиенту f для достижения качества $f(\widehat{x}) - f^* \leqslant \varepsilon$ будет $O\left(\varepsilon^{-2} \log_2 \varepsilon^{-1}\right)$, что близко к оптимальной оценке [36].

Таким образом, верна

Теорема 3.3.9. *Для выхода $\widehat{x}$ модифицированного алгоритма 6 с учетом дополнительной процедуры* (3.14) *неравенство $f(\widehat{x}) - f^* \leqslant \varepsilon$ будет гарантированно выполнено не более, чем после*

$$\left\lceil \frac{4LR^2}{\varepsilon} + \frac{64\Delta^2 R^2}{\varepsilon^2} \right\rceil \cdot \left\lceil \log_2 \left(1 + \frac{16\Delta^2}{\varepsilon L}\right) \right\rceil \tag{3.19}$$

вычислений субградиента f.

Замечание 3.3.10. Отметим, что наблюдаемые экспериментально в данном разделе эффекты о сходимости метода со скоростью $O(\varepsilon^{-1})$ для некоторых выпуклых негладких задач наблюдались и для так называемых универсальных градиентных методов, предложенных недавно Ю. Е. Нестеровым [152]. Однако для негладких задач с липшицевым целевым функционалом доказана была при этом только оценка скорости сходимости вида $O\left(\varepsilon^{-2}\right)$, зависящая ещё от константы Липшица целевого функционала и расстояния от точки старта до ближайшей точки минимума. Найденная нами оценка (3.18) учитывает количество точек негладкости функционала и может быть лучше при малом $\Delta > 0$ (в этом случае доказуемая оценка скорости сходимости близка к $O\left(\varepsilon^{-1}\right)$).

Покажем, что даст применение похожей схемы для ускоренного градиентного метода на классе выпуклых функционалов f, для которых при некоторых $L > 0$ и $\Delta > 0$ верно

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + \Delta\|y - x\| \quad \forall x, y \in Q$$

для некоторого субградиента ∇f. Мы рассматриваем вариант ускоренного градиентного метода (алгоритм 7), который использует (δ, Δ, L)-модель целевого функционала в произвольной запрошенной точке. Заметим, что для его работы уже необходима 1-сильная выпуклость прокс-функции в определении 3.1.1. Применим вариант этого метода при $L_{k+1} = 2^p L$ для фиксированного $p \in \mathbb{N}$ и $\delta = \frac{\varepsilon}{2\gamma}$ для некоторого $\varepsilon > 0$ и постоянной $\gamma > 0$. Будем при этом считать $f_\delta = f$ (то есть функционал f задан точно). Будем подбирать натуральное p далее так, чтобы на $(k+1)$-ой итерации $(k = 0, 1, 2, \ldots)$ гарантированно выполнялось неравенство

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

В таком случае после N итераций (это можно проверить аналогично [22]) будет верно неравенство

$$f(x^N) - f^* \leqslant \frac{8 \cdot 2^p \cdot LR^2}{(N+1)^2} + \frac{\varepsilon N}{2\gamma}.$$

Выберем натуральное p так, чтобы гарантированно выполнялась альтернатива

$$\left[\begin{array}{c} \Delta\|x^{k+1} - y^{k+1}\| \leqslant \frac{\varepsilon}{2\gamma}, \\ \frac{(2^p-1)L}{2}\|x^{k+1} - y^{k+1}\|^2 \geqslant \Delta\|x^{k+1} - y^{k+1}\|. \end{array} \right. \tag{3.20}$$

Если

$$\Delta\|x^{k+1} - y^{k+1}\| > \frac{\varepsilon}{2\gamma},$$

то

$$\frac{(2^p-1)L}{2}\|x^{k+1} - y^{k+1}\| > \frac{(2^p-1)L\varepsilon}{4\gamma\Delta}.$$

Тогда второе неравенство альтернативы (3.20) заведомо выполнится при

$$2^p > 1 + \frac{4\gamma\Delta^2}{L\varepsilon}.$$

Поэтому положим

$$p = \left\lceil \log_2\left(1 + \frac{4\gamma\Delta^2}{L\varepsilon}\right)\right\rceil. \tag{3.21}$$

Теперь покажем, каким можно выбрать количество итераций N, чтобы гарантированно было верно $f(x^N) - f^* \leqslant \varepsilon$. Для этого потребуем выполнения неравенств

$$\frac{2^{p+3}LR^2}{(N+1)^2} \leqslant \frac{\varepsilon}{2} \text{ и } \frac{\varepsilon N}{2\gamma} \leqslant \frac{\varepsilon}{2},$$

откуда $\gamma \geqslant N$ и $(N+1)^2 \geqslant \frac{2^{p+4}LR^2}{\varepsilon}$. Для упрощения выкладок усилим последнее требование: $N^2 \geqslant \frac{2^{p+4}LR^2}{\varepsilon}$, откуда

$$N^2 > \frac{16LR^2}{\varepsilon}\left(1 + \frac{4\gamma\Delta^2}{L\varepsilon}\right) \geqslant \frac{16LR^2}{\varepsilon} + \frac{64N\Delta^2R^2}{\varepsilon^2}.$$

Это означает, что N можно выбирать как $\lceil N_2 \rceil$, где N_2 — больший корень уравнения

$$N^2 - \frac{64\Delta^2R^2}{\varepsilon}N - \frac{16LR^2}{\varepsilon} = 0:$$

$$N_2 = \frac{32\Delta^2R^2}{\varepsilon^2} + \sqrt{\left(\frac{32\Delta^2R^2}{\varepsilon^2}\right)^2 + \frac{16LR^2}{\varepsilon}}.$$

Далее, в силу неравенства $\sqrt{a+b} \geqslant \sqrt{\frac{a}{2}} + \sqrt{\frac{b}{2}}$ $(a, b > 0)$ имеем

$$N \geqslant N_2 \geqslant \frac{(32+16\sqrt{2})\Delta^2R^2}{\varepsilon^2} + \frac{2R\sqrt{2L}}{\sqrt{\varepsilon}}.$$

Таким образом, можно гарантировать достижение ε-точного решения задачи минимизации f за

$$O\left(\sqrt{\frac{L}{\varepsilon}}\right) + O\left(\frac{\Delta^2}{\varepsilon^2}\right)$$

итераций быстрого градиентного метода, т.е. обоснован аналог результата о слайдинге [91].

При этом возможно использовать метод с адаптивной настройкой констант L, а также $\Delta_{k+1} \leqslant \Delta$. Тогда оценка числа итераций может измениться в постоянное число раз. Также при этом за счёт p дополнительных процедур адаптивного подбора L_{k+1} (при условии $L \leqslant L_{k+1} \leqslant 2L$ для всякого $k = 0, 1, \ldots, N-1$) на каждой итерации метода добавится множитель (3.21):

$$p \geqslant \left\lceil \log_2 \left(1 + \frac{4N\Delta^2}{L\varepsilon}\right)\right\rceil \geqslant \log_2 \left(1 + \frac{(128+64\sqrt{2})\Delta^4 R^2}{L\varepsilon^3} + \frac{8R\Delta^2\sqrt{2}}{\sqrt{L\varepsilon^3}}\right).$$

Теорема 3.3.11. *Для выхода x^N модифицированного алгоритма 7 с учетом дополнительной процедуры* (3.14) *при $\Delta_{k+1} = \Delta$ на (k+1)-й итерации ($k = 0, 1, \ldots, N-1$) неравенство $f(x^N) - f^* \leqslant \varepsilon$ будет гарантированно выполнено не более чем после*

$$\left\lceil \frac{(32+16\sqrt{2})\Delta^2 R^2}{\varepsilon^2} + \frac{2R\sqrt{2L}}{\sqrt{\varepsilon}} \right\rceil \cdot \left\lceil \log_2 \left(1 + \frac{(128+64\sqrt{2})\Delta^4 R^2}{L\varepsilon^3} + \frac{8R\Delta^2\sqrt{2}}{\sqrt{L\varepsilon^3}}\right)\right\rceil \tag{3.22}$$

вычислений субградиента f.

Замечание 3.3.12. Если не предполагать, что на (k+1)-й итерации ($k = 0, 1, \ldots, N-1$) модифицированных алгоритмов 6 и 7 выполнено неравенство $L \leqslant L_{k+1} \leqslant 2L$, т.е. предусмотреть полностью адаптивную настройку параметров L и Δ, то оценки (3.18) и (3.22) могут увеличиться не более, чем в

$$\left\lceil \max\left\{\frac{2L}{L_0}, \frac{2\Delta}{\Delta_0}\right\}\right\rceil$$

раз.

Замечание 3.3.13. В начале пункта 3.3 мы говорили, что возможны различные интерпретации параметра Δ. При изложении материала упор был сделан на его смысл как искусственно вводимой в оптимизационную модель неточности для негладких оптимизационных задач. Это сделано потому, что эта интерпретация полученных результатов представляется оригинальной. Однако вполне возможна и более стандартная интерпретация Δ как аддитивной погрешности задания градиента для гладких оптимизационных задач. В этом случае полученные результаты можно рассматривать как вариации известных выводов (см. например, [89]) об отсутствии накопления аддитивной погрешности задания градиента (не целевого функционала) в теоретических оценках скорости сходимости как для неускоренных, так и для ускоренных градиентных методов первого порядка. В таком контексте интерес предложенных подходов и доказанных результатов определяется адаптивностью рассматриваемых методов (в том числе и относительно параметра Δ), а также самими оценками в виде распределения сложностей (что в некотором смысле похоже на результаты о градиентном слайдинге [120]).

В качестве примера негладкой задачи с искусственной неточностью рассмотрим расчёты для аналога задачи Ферма–Торричелли–Штейнера.

Пример 3.3.14. Для случайно подобранных точек $A_k = (a_{1k}, a_{2k}, \ldots, a_{nk})$ (A_k таковы, что $1 < \sqrt{a_{1k}^2 + a_{2k}^2 + \ldots + a_{nk}^2} < 10$, $k = \overline{1,m}$, $m = 50$) в единичном шаре n-мерного евклидова пространства $\mathbb{R}^n$ ($n = 10^5$) рассмотрим задачу нахождения точки $X = (x_1, x_2, \ldots, x_n)$, доставляющей минимум функционалу $f(x) := \sum\limits_{k=1}^{m} XA_k$ на шаре радиуса $R = 10$ с центром в нуле, где

$$XA_k = \sqrt{(x_1 - a_{1k})^2 + (x_2 - a_{2k})^2 + \ldots + (x_n - a_{nk})^2}.$$

Поскольку данные сгенерированы случайно, то качество решения разумно оценивать именно по теоретическим оценкам. В таблицах 3.5 и 3.6 приведены усреднённые результаты 10 экспериментов с различным выбором начальных точек в зависимости от количества итераций за сопоставимое время. Как видим, адаптивность приводит к достижению существенно лучшего качества решения. При этом адаптивный

Таблица 3.5. Результаты для задачи Ферма–Торричелли–Штейнера для предложенного варианта неускоренного градиентного метода.

K	Адаптивный		Неадаптивный	
	Оценка	Время, с	Оценка	Время, с
10	13.4714	9.2	12.8383	9.2
20	6.2867	12.9	7.3916	13.0
30	4.1000	16.3	5.6365	16.8
40	3.04200	19.9	4.7249	20.2
50	2.4180	23.6	4.1569	24.0
100	1.1937	41.7	3.0193	41.9
200	0.5931	77.9	2.4397	78.7
300	0.3946	113.7	2.2484	115.8

ускоренный метод работает эффективнее по сравнению с адаптивным неускоренным, несмотря на теоретическую возможность накопления величин, соответствующих погрешностям. Возможность накопления погрешностей как раз видна, если сравнить результаты для неадаптивных методов.

3.4 Методы для вариационных неравенств с адаптивной настройкой на величины погрешностей

Теперь рассмотрим концепции неточной оптимизационной модели с двумя параметрами погрешностей уже для вариационных неравенств и седловых задач. На выделенном классе задач метод с адаптивной настройкой на величину параметра гладкости L и параметров погрешностей. Начнём с частного случая вариационных неравенств с оператором G, который удовлетворяет на некотором выпуклом компакте $Q \subset \mathbb{R}^n$ следующему аналогу условия Липшица:

$$\langle G(y) - G(x), y - z\rangle \leqslant LV(y,x) + LV(z,y) + \Delta||y - x||$$

для фиксированной величины $\Delta > 0$.

Рассмотрим поясняющий пример.

Таблица 3.6. Результаты экспериментов для задачи Ферма–Торричелли–Штейнера для предложенного варианта быстрого градиентного метода.

K	Адаптивный		Неадаптивный	
	Оценка	Время, с	Оценка	Время, с
10	0.5083	2.1	13.1725	3.0
20	0.1423	6.9	21.2397	7.2
30	0.0686	11.0	27.3032	10.6
40	0.0428	14.5	30.8354	14.3
50	0.0316	18.1	33.5755	17.9
100	0.0241	36.8	38.5851	35.9
200	0.0407	73.4	39.6018	72.3
300	0.0594	110.3	39.6330	108.0

Пример 3.4.1. Пусть $G : Q \to \mathbb{R}^n$ — липшицев оператор с константой Липшица $L > 0$, т.е.

$$\|G(x) - G(y)\|_* \leqslant L\|x - y\| \quad \forall\, x,\, y \in Q.$$

Однако предположим, что точное значение оператора G недоступно и известно только приближённое его значение, такое, что

$$\|\widetilde{G}(x) - G(x)\|_* \leqslant \frac{\Delta}{2} \quad \forall\, x \in Q.$$

Тогда для всяких x, y, $z \in Q$:

$$\begin{aligned}
&|\langle \widetilde{G}(y) - \widetilde{G}(x),\, y - z\rangle - \langle G(y) - G(x),\, y - z\rangle| = \\
&= |\langle \widetilde{G}(y) - G(y),\, y - z\rangle + \langle G(x) - \widetilde{G}(x),\, y - z\rangle| \leqslant \\
&\leqslant \|\widetilde{G}(y) - G(y)\|_* \cdot \|y - z\| + \|\widetilde{G}(x) - G(x)\|_* \cdot \|y - z\| \leqslant \\
&\leqslant \left(\frac{\Delta}{2} + \frac{\Delta}{2}\right)\|y - z\| = \Delta\|y - z\|.
\end{aligned}$$

Следовательно,

$$\begin{aligned}
&\langle \widetilde{G}(y) - \widetilde{G}(x),\, y - z\rangle \leqslant \langle G(y) - G(x),\, y - z\rangle + \Delta\|y - z\| \leqslant \\
&\leqslant \|G(y) - G(x)\|_* \cdot \|y - z\| + \Delta\|y - z\| \leqslant \\
&\leqslant L\|y - x\| \cdot \|y - z\| + \Delta\|y - z\| \leqslant \frac{L}{2}\|y - x\|^2 + \frac{L}{2}\|y - z\|^2 + \Delta\|y - z\| \leqslant \\
&\leqslant LV(y, x) + LV(y, z) + \Delta.
\end{aligned}$$

$$\langle G(y)-G(x),\, y-z\rangle \leqslant LV(y,x)+LV(y,z)+\Delta\|y-z\| \quad \forall x,y,z\in Q, \tag{3.23}$$

Замечание 3.4.2. Отметим, что слагаемое $\Delta||y-x||$ описывает степень отклонения от непрерывности для оператора G вдоль всякого фиксированного векторного отрезка $[x;y] = ty + (1-t)x$ при $0 \leqslant t \leqslant 1$. В целом (если объединить все возможные векторные отрезки) на области определения точек разрыва оператора G может быть бесконечно много.

Пример 3.4.3. Обратим внимание, что выражение $\Delta\|y-z\|$ в (3.23) может описывать отклонения от непрерывности для оператора G вдоль любого фиксированного векторного отрезка $\{ty+(1-t)x\}_{0\leqslant t\leqslant 1}$. В общем случае (если объединить все возможные векторные сегменты) в области точек негладкости может быть бесконечное число. Например, предположим, что для некоторого подмножества $Q_0 \subset Q$ функция f дифференцируема во всех точках $Q\backslash Q_0$ и что для произвольного $x \in Q_0$ существует конечный субдифференциал $\partial f(x)$ в смысле выпуклого анализа. Для фиксированных $x, y \in Q$ с $t \in [0;1]$ обозначается как $y_t := (1-t)x + ty$. Пусть выпуклая функция $f : Q \to \mathbb{R}$ ($Q \subset \mathbb{R}^n$) имеет (Δ, L)- липшицев субградиент ($f \in C^{1,1}_{L,\Delta}(Q)$), которое было введено ранее. Это означает, что для всякого субградиента $G(x) = \widehat{\partial} f(x)$ верно

$$\|G(y) - G(x)\|_* \leqslant L\|y-x\| + \Delta. \tag{3.24}$$

Чтобы доказать (3.24), достаточно разбить отрезок $\{y_t\}_{0\leqslant t\leqslant 1}$ на интервалы гладкости и учесть ограниченность диаметров субдифференциалов при негладкости точки f.

Неравенство (3.24) означает, что

$$\langle G(y)-G(x),\, y-z\rangle \leqslant \|G(y)-G(x)\|_*\cdot\|y-z\| \leqslant L\|y-x\|\cdot\|y-z\|+\Delta\|y-z\| \leqslant$$

$$\leqslant \frac{L}{2}(\|y-x\|^2 + \|y-z\|^2) + \Delta\|y-z\| \leqslant LV(y,x) + LV(z,y) + \Delta\|y-z\|.$$

Любой функционал f с конечным набором негладких точек вдоль произвольного отрезка будет удовлетворять предлагаемому условию Липшица для субградиента. Очевидно, что это условие выполняется для каждой целевой функции с конечными точками негладкости на каждом векторном отрезке $[x;y]$. Таким образом, эту технику можно применить, например, к задачам минимизации суммы расстояний до

нескольких шаров в гильбертовых пространствах [131]. Такая целевая функция, очевидно, не будет дифференцируемой в обычном смысле в точках границ шаров, которых бесконечно много. Отметим, что среди точек каждого векторного отрезка $[x; y]$ такая целевая функция имеет конечное число точек негладкости. Однако рассмотренное условие Липшица при специальном выборе субградиента выполняется для некоторых функций и с бесконечным числом точек негладкости (можно выбрать $\Delta = 2M$, если f M-липшицев).

3.4.1 Аналог проксимального зеркального метода для вариационных неравенств с адаптацией к величинам погрешностей

В этом разделе мы рассмотрим новую версию проксимального зеркального метода для вариационных неравенств (*проксимальный зеркальный метод с адаптацией к неточности (ПЗМАН)*). Для данного метода рассмотрена адаптация не только к уровню гладкости оператора, но также к величине ошибки оракула, которая может позволить получить трудоёмкость около $O\left(\frac{1}{\varepsilon}\right)$ для вариационных неравенств с ограниченными операторами, что близко к оптимальной оценке на существенно более узком классе вариационных неравенств с липшицевыми операторами.

Теорема 3.4.4. *После выполнения N итераций алгоритма 8, верно следующее неравенство:*

$$\sum_{k=0}^{N-1} \frac{1}{L_{k+1}} \langle G(y^{k+1}),\, y^{k+1} - x\rangle \leqslant$$

$$\leqslant V(x, x^0) - V(x, x^N) + \sum_{k=0}^{N-1} \frac{\Delta_{k+1}}{L_{k+1}} \left\|y^{k+1} - x^{k+1}\right\|.$$

Доказательство. Непосредственно проверяются следующие соотношения:

$$\left\langle \nabla_x V(x, x^k)\big|_{x=x^{k+1}},\, x - x^{k+1}\right\rangle = V(x, x^k) - V(x, x^{k+1}) - V(x^{k+1}, x^k),$$

$$\left\langle \nabla_x V(x, x^k)\big|_{x=y^{k+1}},\, x - y^{k+1}\right\rangle = V(x, x^k) - V(x, y^{k+1}) - V(y^{k+1}, x^k).$$

Далее, для всякого $x \in Q$ и $k = \overline{0,\ N-1}$:

$$\left\langle \nabla_x \left(\left\langle G(x^k),\, x - x^k\right\rangle + L_{k+1} V(x, x^k)\right)\big|_{x=y^{k+1}},\ x - y^{k+1}\right\rangle \geqslant 0,$$

Алгоритм 8 Проксимальный зеркальный метод с адаптацией к величинам погрешностей.

Input: $x^0 = \arg\min\limits_{x \in Q} d(x), L_0, \Delta_0$.

1: $N := N + 1$; $L_{N+1} := \frac{L_N}{2}$; $\Delta_{N+1} := \frac{\Delta_N}{2}$.

2: Вычислить

$$y^{N+1} := \arg\min_{x \in Q}\{\langle G(x^N),\, x - x^N\rangle + L_{N+1}V(x, x^N)\}, \tag{3.25}$$

$$x^{N+1} := \arg\min_{x \in Q}\{\langle G(y^{N+1}),\, x - x^N\rangle + L_{N+1}V(x, x^N)\}. \tag{3.26}$$

3: **If**

$$\langle G(y^{N+1}) - G(x^N),\, y^{N+1} - x^{N+1}\rangle \leqslant L_{N+1}V(y^{N+1}, x^N) + \tag{3.27}$$

$$+L_{N+1}V(x^{N+1}, y^{N+1}) + \Delta_{N+1}\left\|y^{N+1} - x^{N+1}\right\|,$$

then переходим к следующей итерации (пункт 1).

4: **Else** $L_{N+1} := 2L_{N+1}$, $\Delta_{N+1} := 2\Delta_{N+1}$ и переходим к пункту 2.

$$\left\langle \nabla_x \left(\left\langle G(y^{k+1}),\, x - x^k\right\rangle + L_{k+1}V(x, x^k)\right)\Big|_{x=x^{k+1}},\; x - x^{k+1}\right\rangle \geqslant 0.$$

Поэтому верно неравенство

$$\left\langle G(y^{k+1}),\, x^{k+1} - x\right\rangle \leqslant L_{k+1}V(x, x^k) - L_{k+1}V(x, x^{k+1}) - L_{k+1}V(x^{k+1}, x^k)$$

а также

$$\left\langle G(x^k),\, y^{k+1} - x\right\rangle \leqslant L_{k+1}V(x, x^k) - L_{k+1}V(x, y^{k+1}) - L_{k+1}V(y^{k+1}, x^k).$$

С учётом (3.27) для всякого $k = \overline{0,\; N-1}$ имеем:

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

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

$$\leqslant L_{k+1}V(x, x^k) - L_{k+1}V(x, x^{k+1}) - L_{k+1}V(x^{k+1}, x^k) + L_{k+1}V(x^{k+1}, x^k) -$$

$$-L_{k+1}V(x^{k+1}, y^{k+1}) - L_{k+1}V(y^{k+1}, x^k) + L_{k+1}V(y^{k+1}, x^k) +$$

$$+L_{k+1}V(x^{k+1}, y^{k+1}) + \sum_{k=0}^{N-1} \frac{\Delta_{k+1}}{L_{k+1}}\left\|y^{k+1} - x^{k+1}\right\|,$$

т.е.

$$\frac{1}{L_{k+1}}\left\langle G(y^{k+1}),\, y^{k+1}-x\right\rangle \leqslant$$
$$\leqslant V(x,x^k)-V(x,x^{k+1})+\sum_{k=0}^{N-1}\frac{\Delta_{k+1}}{L_{k+1}}\left\|y^{k+1}-x^{k+1}\right\|. \tag{3.28}$$

После суммирования неравенств вида (3.28) по $k=\overline{0,\ N-1}$ имеем

$$\sum_{k=0}^{N-1}\frac{1}{L_{k+1}}\langle G(y^{k+1}),\, y^{k+1}-x\rangle \leqslant$$
$$\leqslant V(x,x^0)-V(x,x^N)+\sum_{k=0}^{N-1}\frac{\Delta_{k+1}}{L_{k+1}}\left\|y^{k+1}-x^{k+1}\right\|.$$

□

Введем обозначение

$$S_N=\sum_{k=0}^{N-1}\frac{1}{L_{k+1}},\ \widetilde{y}=\frac{1}{S_N}\sum_{k=0}^{N-1}\frac{y^{k+1}}{L_{k+1}}\ \text{и}\ R^2=\max_{x\in Q}V(x,x^0).$$

Теорема 3.4.5. *Для монотонного оператора G после N итераций алгоритма 8 справедлива следующая оценка:*

$$\max_{x\in Q}\langle G(x),\,\widetilde{y}-x\rangle \leqslant \frac{R^2}{S_N}+\frac{1}{S_N}\sum_{k=0}^{N-1}\frac{\Delta_{k+1}}{L_{k+1}}\left\|y^{k+1}-x^{k+1}\right\|. \tag{3.29}$$

Предположим, что для фиксированного ε

$$\sum_{k=0}^{N-1}\frac{1}{L_{k+1}}\geqslant\frac{R^2}{\varepsilon}. \tag{3.30}$$

Тогда имеет место следующее неравенство:

$$\max_{x\in Q}\langle G(x),\,\widetilde{y}-x\rangle \leqslant \varepsilon+\frac{1}{S_N}\sum_{k=0}^{N-1}\frac{\Delta_{k+1}}{L_{k+1}}\left\|y^{k+1}-x^{k+1}\right\|.$$

Если $L_0\leqslant 2L$, тогда неравенство (3.30) выполняется не более, чем после

$$N=\left\lceil\frac{2LR^2}{\varepsilon}\right\rceil$$

итераций алгоритма 8.

Доказательство. Ввиду монотонности оператора G для всякого $k = 0, 1, ...$ имеем:

$$\langle G(x),\, y^{k+1} - x\rangle = \langle G(y^{k+1}),\, y^{k+1} - x\rangle + \langle G(x) - G(y^{k+1}),\, y^{k+1} - x\rangle \leqslant$$

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

Поэтому неравенство

$$\frac{1}{S_N} \max_{x\in Q} \sum_{k=0}^{N-1} \frac{1}{L_{k+1}} \left\langle G(y^{k+1}),\, y^{k+1} - x\right\rangle \leqslant$$

$$\leqslant \frac{R^2}{S_N} + \frac{1}{S_N} \sum_{k=0}^{N-1} \frac{\Delta_{k+1}}{L_{k+1}} \left\|y^{k+1} - x^{k+1}\right\| \leqslant \varepsilon + \frac{1}{S_N} \sum_{k=0}^{N-1} \frac{\Delta_{k+1}}{L_{k+1}} \left\|y^{k+1} - x^{k+1}\right\|$$

можно заменить на

$$\begin{aligned} \max_{x\in Q} \langle G(x),\, \widetilde{y} - x\rangle &\leqslant \frac{R^2}{S_N} + \frac{1}{S_N} \sum_{k=0}^{N-1} \frac{\Delta_{k+1}}{L_{k+1}} \left\|y^{k+1} - x^{k+1}\right\| \leqslant \\ &\leqslant \varepsilon + \frac{1}{S_N} \sum_{k=0}^{N-1} \frac{\Delta_{k+1}}{L_{k+1}} \left\|y^{k+1} - x^{k+1}\right\|. \end{aligned} \tag{3.31}$$

□

Замечание 3.4.6. В силу адаптивного выбора параметров L_{k+1} и Δ_{k+1} на каждой итерации алгоритма 8 выражение

$$\frac{R^2}{S_N} + \frac{1}{S_N} \sum_{k=0}^{N-1} \frac{\Delta_{k+1}}{L_{k+1}} \left\|y^{k+1} - x^{k+1}\right\|$$

в (3.31) может быть достаточно малым даже в случае $L = +\infty$ или $\Delta = +\infty$ в (3.23).

Замечание 3.4.7. Очевидно, что для всякого k верно $\Delta_k \leqslant C_L \Delta$ ($C_L = \max\left\{1, \frac{2L}{L_0}\right\}$) и:

$$\frac{1}{S_N} \sum_{k=0}^{N-1} \frac{\Delta_{k+1}}{L_{k+1}} \left\|y^{k+1} - x^{k+1}\right\| \leqslant C_L \Delta \max_{k=0,N-1} \left\|y^{k+1} - x^{k+1}\right\|.$$

Это означает, что величина, связанная с неточностью задания оператора G, ограничена на множестве Q конечного диаметра.

Замечание 3.4.8. Если $G \not\equiv 0$, то условие $L_0 \leqslant 2L$ будет заведомо верно, если

$$L_0 := \frac{\|G(x) - G(y)\|_*}{\|x - y\|} \text{ если } G(x) \neq G(y)$$

для некоторых x и y.

Замечание 3.4.9. В общем случае, если рассматривать задачи решения вариационных неравенств с погрешностями задания монотонного оператора G, то его приближение (оператор $\widetilde{G}$) может потерять свойство монотонности. Так, если для некоторого $\Delta > 0$ верно $\left\|\widetilde{G}(x) - G(x)\right\|_* \leqslant \frac{\Delta}{2}\ \forall x \in Q$, то можно лишь утверждать

$$\left\langle \widetilde{G}(x) - \widetilde{G}(y), x - y \right\rangle \geqslant \langle G(x) - G(y), x - y\rangle - \Delta \|x - y\| \geqslant -\Delta \|x - y\|$$

для произвольных $x, y \in Q$. В таком случае, если применить алгоритм 8 для вариационного неравенства с оператором $\widetilde{G}$, то после остановки метода будет выполнен следующий аналог (3.29):

$$\max_{x \in Q} \left\langle \widetilde{G}(x),\, \widetilde{y} - x \right\rangle \leqslant \frac{R^2}{S_N} + \frac{1}{S_N} \sum_{k=0}^{N-1} \frac{\Delta_{k+1}}{L_{k+1}} \left\|y^{k+1} - x^{k+1}\right\| + \Delta \max_{x \in Q} \|\widetilde{y} - x\|\,. \tag{3.32}$$

Замечание 3.4.10. Заметим, что оценка количества итераций $N = \left\lceil \frac{2LR^2}{\varepsilon} \right\rceil$ в зависимости от желаемой точности ε решения задачи вариационного неравенства оптимальна с точностью до константы в случае липшицева оператора. [134]. При этом величина погрешности, как можно видеть из предыдущего замечания, конечна и не накапливается. При этом можно проверить, что общее количество обращений на итерациях метода ко вспомогательным задачам вида (3.25) и (3.26) ограничено числом

$$4N + \max\left\{\log_2 \frac{2L}{L_0},\ \log_2 \frac{2\Delta}{\Delta_0}\right\}.$$

Замечание 3.4.11. Как отмечено ранее, величина Δ может в некотором смысле выступать характеристикой скачков значений в точках разрыва оператора G (отклонение от непрерывности). Покажем, как возможно немного модифицировать алгоритм 8, чтобы достичь качество решения $\widetilde{y}$:

$$\max_{x \in Q} \langle G(x),\, \widetilde{y} - x\rangle \leqslant \varepsilon$$

за

$$O\left(\frac{1}{\varepsilon^2}\log_2\frac{1}{\varepsilon}\right)$$

итераций для ограниченного оператора G (вообще говоря, негладкого).

Предположим, что на каждой итерации для L_{k+1} $L \leqslant L_{k+1} \leqslant 2L$ (этого всегда можно достичь за несколько шагов п. 2 листинга алгоритма 8).

Предложим следующую процедуру для некоторого натурального p: повторять операцию пункта 2 алгоритма 8, увеличивая L_{k+1} при сохранении $\Delta_{k+1} \leqslant 2\Delta$:

$$L_{k+1} := 2 \cdot L_{k+1} \text{ при сохранении } \Delta_{k+1} \leqslant 2\Delta \tag{3.33}$$

(на каждой итерации алгоритма 8, $k = 0, 1, 2, \ldots, N-1$). Процедуру (3.33) остановим, когда будет верно одно из неравенств:

$$\Delta_{k+1}\left\|y^{k+1} - x^{k+1}\right\| \leqslant \frac{\varepsilon}{2}, \tag{3.34}$$

или

$$\begin{gathered}\left\langle G(y^{k+1}) - G(x^k),\, y^{k+1} - x^{k+1}\right\rangle \leqslant \\ \leqslant 2^{p-1}L\left(\left\|y^{k+1} - x^k\right\|^2 + \left\|y^{k+1} - x^{k+1}\right\|^2\right).\end{gathered} \tag{3.35}$$

Ясно, что (3.33) предполагает обновление x^{k+1} и y^{k+1}. Оценим необходимое количество повторенной для достижения (3.34) или (3.35).

Ясно, что для произвольных $x^k, x^{k+1}, y^{k+1} \in Q$ верно

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

$$\leqslant \frac{L}{2}\left\|y^{k+1} - x^k\right\|^2 + \frac{L}{2}\left\|y^{k+1} - x^{k+1}\right\|^2 + \Delta\left\|y^{k+1} - x^{k+1}\right\|.$$

Более того, $\Delta_{k+1} \leqslant 2\Delta$. Если (3.34) не верно, то $\left\|y^{k+1} - x^{k+1}\right\| > \dfrac{\varepsilon}{4\Delta}$ и неравенство (3.35) гарантированно выполнено при

$$2^p > 1 + \frac{16\Delta^2}{L\varepsilon}, \tag{3.36}$$

поскольку в таком случае

$$\frac{2^p - 1}{2}L\left\|y^{k+1} - x^{k+1}\right\|^2 > \Delta\left\|y^{k+1} - x^{k+1}\right\|.$$

Поэтому после повторения p процедур типа (3.33) на каждой из N итераций базового алгоритма будет выполнено неравенство:

$$\max_{x\in Q}\langle G(x),\, \widetilde{y}-x\rangle \leqslant \frac{R^2}{S_N}+\frac{\varepsilon}{2}.$$

Далее,

$$S_N=\sum_{k=0}^{N-1}\frac{1}{L_{k+1}}\geqslant\frac{N}{2^{p+1}L}.$$

Поэтому $\dfrac{R^2}{S_N}\leqslant\dfrac{2^{p+1}LR^2}{N}\leqslant\dfrac{\varepsilon}{2}$ при $N\geqslant\dfrac{2^{p+2}LR^2}{\varepsilon}$, откуда с учетом (3.36)

$$N\geqslant\frac{4LR^2}{\varepsilon}+\frac{64\Delta^2R^2}{\varepsilon^2}.$$

Вообще говоря, необходимо $O\left(\log\dfrac{1}{\varepsilon}\right)$ дополнительных шагов пункта 2 алгоритма 8 на каждой итерации. Таким образом, для достижения качества решения $\max\limits_{x\in Q}\langle G(x),\, \widetilde{y}-x\rangle\leqslant\varepsilon$ достаточно $O\left(\dfrac{1}{\varepsilon^2}\log_2\dfrac{1}{\varepsilon}\right)$ обращений к оракулу для оператора поля. Известно, что эта оценка оптимальна с точностью до логарифмического множителя.

3.4.2 Численные эксперименты: билинейные матричные игры с погрешностью

Далее рассмотрим расчеты для задачи нахождения равновесия по Нэшу в матричной игре. Рассмотрим седловую задачу

$$\min_{x\,\in\,\Delta_n}\max_{y\,\in\,\Delta_m} x^TAy, \tag{3.37}$$

где $x=(x_1,\,x_2,\,\ldots,\,x_n)\in\mathbb{R}^n$, $y=(y_1,\,y_2,\,\ldots,\,y_m)\in\mathbb{R}^m$, Δ_n — единичный симплекс в $\mathbb{R}^n$, то есть $\Delta_n=\{x\in\mathbb{R}^n\,|\,x\geq 0,\,\sum\limits_{i=1}^{n}x_i=1\}$, Δ_m — единичный симплекс в $\mathbb{R}^m$, A это матрица выплат для игрока y, размер матрицы $m\times n$. Рассмотрим следующий оператор

$$G(u)=\begin{pmatrix}\nabla_x(x^TAy)\\ -\nabla_y(x^TAy)\end{pmatrix}=\begin{pmatrix}A^Ty\\ -Ax\end{pmatrix},\ u=(x,y)\in Q\equiv\Delta_n\times\Delta_m,$$

Предположим, что нам доступно некоторое приближение оператора G

$$\widetilde{G}(u) = \begin{pmatrix} \nabla_x(x^T A y) + \frac{\widetilde{\delta}}{2n} \\ -\nabla_y(x^T A y) + \frac{\widetilde{\delta}}{2m} \end{pmatrix} = \begin{pmatrix} A^T y + \frac{\widetilde{\delta}}{2n} \\ -Ax + \frac{\widetilde{\delta}}{2m} \end{pmatrix}, \quad u = (x, y) \in Q \equiv \Delta_n \times \Delta_m, \tag{3.38}$$

где $\widetilde{\delta} \in \left(-\frac{1}{3}, \frac{1}{3}\right)$. Оператор $\widetilde{G}(u)$ из (3.38) монотонен на Q. Поэтому для такого оператора вариационное неравенство

$$\langle G(x),\, x_* - x\rangle \leqslant 0 \;\; \forall\, x \in Q.$$

имеет решение, совпадающее с решением седловой задачи (3.37).

Поэтому для указанной задачи возможно использовать методы как с адаптивной настройкой на величину погрешности (алгоритм 8 с оценкой (3.32)), так и с фиксированной величиной Δ, соответствующей погрешности задания оператора (без адаптивной настройки). По аналогии с пунктом 2.4.5 рассматривается энтропийная прокс-функция вида $d(x) = \sum\limits_{i=1}^{n+m} x_i \ln x_i$ и соответствующая ей дивергенция (расхождение Кульбака–Лейбнера).

Сравним результаты выполнения алгоритма 8 с оценкой (3.32) и вариантом этого метода с фиксированной величиной $\Delta_{k+1} = \Delta$ ($k \in 0, 1, \ldots$). Как видим, использование дополнительной адаптивной настройки на параметры неточности оптимизационной модели позволяет за сопоставимое время гарантировать лучшее качество решения примерно в два раза. Отметим при этом, что оба метода адаптивны по величине константы гладкости L. Усредненные результаты выполнения указанных методов для 5 экспериментов представлены в таблице 3.7, матрица A составлена случайным образом из целых чисел от 0 до 9.

3.4.3 Адаптивный метод для вариационных неравенств и седловых задач, допускающих существование в произвольной точке (δ, Δ, L)-модели

В этом пункте мы покажем, как можно предложить аналог методики пункта 3.4.1 в модельной общности, а также с использованием двух параметров погрешностей.

Напомним постановку задачи решения вариационного неравенства в модельной общности, а также необходимые понятия и результаты. Для

Таблица 3.7. Результаты работы алгоритма, $m = 1000,\ n = 2000$.

N	Адаптивный		Неадаптивный	
	Оценка	Время, с	Оценка	Время, с
10	0.48364	5.8	0.86992	4.4
20	0.41483	11	0.70761	9.2
50	0.39322	28	0.67267	26
100	0.38945	56.8	0.66903	55.4
200	0.38987	112.6	0.66778	112
300	0.38908	168.4	0.66732	167.6
400	0.38895	224.6	0.66713	223
500	0.38891	279.8	0.66702	278.6
1000	0.38889	565.4	0.66688	554.6

оператора $G : Q \to \mathbb{R}^n$, заданного на выпуклом компакте $Q \subset \mathbb{R}^n$ под *вариационным неравенством* понимаем неравенство вида

$$\langle G(x_*),\ x_* - x\rangle \leqslant 0. \tag{3.39}$$

Отметим, что в (3.39) требуется найти $x_* \in Q$ (это x_* и называется (строгим) решением ВН), для которого

$$\max_{x\in Q}\langle G(x_*),\ x_* - x\rangle \leqslant 0.$$

Для монотонного оператора поля G можно рассматривать *задачу отыскания слабого решения вариационного неравенства*

$$\langle G(x),\ x_* - x\rangle \leqslant 0, \tag{3.40}$$

то есть нахождения $x_* \in Q$, такого, что (3.40) верно при всех $x \in Q$.

Введём аналог понятия неточной оптимизационной (δ, Δ, L)-модели целевой функции для вариационных неравенств и седловых задач. Для удобства будем рассматривать задачу нахождения решения $x_* \in Q$ абстрактной задачи равновесного программирования

$$\psi(x, x_*) \geqslant 0 \quad \forall x \in Q \tag{3.41}$$

для некоторого выпуклого компакта $Q \subset \mathbb{R}^n$, а также функционала $\psi : Q \times Q \to \mathbb{R}$. Если предположить абстрактную монотонность функционала ψ:

$$\psi(x, y) + \psi(y, x) \leqslant 0\ \ \forall x, y \in Q,$$

то всякое решение (3.41) будет также и решением двойственной задачи равновесного программирования

$$\psi(x_*, x) \leqslant 0 \quad \forall x \in Q. \tag{3.42}$$

В общем случае сделаем предположение о существовании решения x_* задачи (3.41). При указанных обозначениях понятие (δ, Δ, L)-модели для выделенного выше класса задач возможно ввести следующим образом.

Определение 3.4.12. Будем говорить, что функционал ψ допускает (δ, Δ, L)-модель $\psi_\delta(x, y)$ при некоторых фиксированных значениях параметров $L,\ \delta,\ \Delta > 0$ на Q относительно дивергенции Брегмана $V(y, x)$, если для произвольных $x, y, z \in Q$ верны:

(i) $\psi(x, y) \leqslant \psi_\delta(x, y) + \delta$;

(ii) $\psi_\delta(x, y)$ — выпуклый функционал по первой переменной и $\psi_\delta(x, x) = 0$;

(iii) (*абстрактная δ-монотонность*)

$$\psi_\delta(x, y) + \psi_\delta(y, x) \leqslant \delta;$$

(iv) (*обобщённая относительная гладкость*)

$$\psi_\delta(x, y) \leqslant \psi_\delta(x, z) + \psi_\delta(z, y) + LV(x, z) + LV(z, y) + \delta + \Delta \|y - z\|. \tag{3.43}$$

Естественно возникает идея обобщить этот метод на абстрактные задачи (3.41) и (3.42) в предположениях их разрешимости, а также (i)–(iv). При этом будем учитывать погрешность δ в (3.43), а также погрешность $\widetilde{\delta}$ решения вспомогательных задач на итерациях согласно одному из достаточно известных в алгоритмической оптимизации подходов:

$$x := \overset{\widetilde{\delta}}{\arg\min_{y \in Q}} \varphi(y), \text{ если } \langle \nabla\varphi(x), x - y\rangle \leqslant \widetilde{\delta} \quad \forall y \in Q.$$

Погрешность $\widetilde{\delta}$ решения вспомогательных подзадач важно учитывать, поскольку речь идет о функциях-моделях общего вида.

К задачам вида (3.41) и (3.42) при условии существования (δ, Δ, L)-модели возможно применить аналог проксимального зеркального метода с адаптивным выбором шага. В отличие от алгоритма 1 здесь предполагается адаптивная настройка не только на L, но и на оба параметра погрешностей. Опишем $(N+1)$-ую итерацию предлагаемого метода $(N = 0, 1, 2, \ldots)$, выбрав начальное приближение $x^0 = \arg\min\limits_{x \in Q} d(x)$, зафиксировав точность $\varepsilon > 0$, а также некоторые константы $L_0 \leqslant 2L$, $\delta_0 \leqslant 2\delta$ и $\Delta_0 \leqslant 2\Delta$.

Алгоритм 9 Адаптивный метод для ВН, допускающих (δ, Δ, L)-модель.

---

1. $N := N + 1$, $L_{N+1} := \frac{L_N}{2}$, $\delta_{N+1} := \frac{\delta_N}{2}$, $\Delta_{N+1} := \frac{\Delta_N}{2}$.

2. Вычисляем:
$$y^{N+1} := \arg\min_{x \in Q}^{\widetilde{\delta}} \left\{\psi_\delta(x, x^N) + L_{N+1} V(x, x^N)\right\},$$
$$x^{N+1} := \arg\min_{x \in Q}^{\widetilde{\delta}} \left\{\psi_\delta(x, y^{N+1}) + L_{N+1} V(x, x^N)\right\}$$
до тех пор, пока не будет выполнено:
$$\begin{aligned} \psi_\delta(x^{N+1}, x^N) \leqslant \\ \leqslant \psi_\delta(y^{N+1}, x^N) + \psi_\delta(x^{N+1}, y^{N+1}) + \\ + L_{N+1} V(y^{N+1}, x^N) + L_{N+1} V(y^{N+1}, x^{N+1}) + \\ + \Delta_{N+1} \left\|y^{N+1} - x^{N+1}\right\| + \delta_{N+1}. \end{aligned} \tag{3.44}$$

3. **Если** (3.44) не выполнено, **то** $L_{N+1} := 2L_{N+1}$, $\delta_{N+1} := 2\delta_{N+1}$, $\Delta_{N+1} := 2\Delta_{N+1}$ и повторяем п. 2.

4. **Иначе** переход к п. 1.

5. Критерий остановки метода:
$$S_N := \sum_{k=0}^{N-1} \frac{1}{L_{k+1}} \geqslant \frac{\max\limits_{x \in Q} V(x, x^0)}{\varepsilon}.$$

Справедлива следующая

Теорема 3.4.13. *После остановки рассматриваемого метода для всякого $x \in Q$ будет заведомо выполнено неравенство:*

$$-\frac{1}{S_N}\sum_{k=0}^{N-1}\frac{\psi_\delta(x, y^{k+1})}{L_{k+1}} \leqslant \varepsilon + 2\widetilde{\delta} + \frac{1}{S_N}\sum_{k=0}^{N-1}\frac{\delta_{k+1} + \Delta_{k+1}\left\|y^{k+1} - x^{k+1}\right\|}{L_{k+1}},$$

а также

$$\psi(\widetilde{y}, x) \leqslant \varepsilon + 2\widetilde{\delta} + 2\delta + \frac{1}{S_N}\sum_{k=0}^{N-1}\frac{\delta_{k+1} + \Delta_{k+1}\left\|y^{k+1} - x^{k+1}\right\|}{L_{k+1}} \tag{3.45}$$

при

$$\widetilde{y} := \frac{1}{S_N}\sum_{k=0}^{N-1}\frac{y^{k+1}}{L_{k+1}}.$$

Для обычных слабых вариационных неравенств (3.40) неравенство (3.45) можно заменить на

$$\max_{x\in Q}\langle G(x), \widetilde{y} - x\rangle \leqslant \varepsilon + 2\widetilde{\delta} + 3\delta + \frac{1}{S_N}\sum_{k=0}^{N-1}\frac{\delta_{k+1} + \Delta_{k+1}\left\|y^{k+1} - x^{k+1}\right\|}{L_{k+1}}.$$

Можно ввести также аналог понятия (δ, Δ, L)-модели для седловых задач. Напомним, что постановка седловой задачи предполагает, что для выпуклого по u и вогнутого по v функционала $f(u, v) : \mathbb{R}^{n_1+n_2} \to \mathbb{R}$ ($u \in Q_1 \subset \mathbb{R}^{n_1}$ и $v \in Q_2 \subset \mathbb{R}^{n_2}$) требуется найти (u_*, v_*) такую, что:

$$f(u_*, v) \leqslant f(u_*, v_*) \leqslant f(u, v_*) \tag{3.46}$$

для произвольных $u \in Q_1$ и $v \in Q_2$. Пусть Q_1 и Q_2 — выпуклые компакты в пространствах $\mathbb{R}^{n_1}$ и $\mathbb{R}^{n_2}$ и поэтому $Q = Q_1 \times Q_2 \subset \mathbb{R}^{n_1+n_2}$ также есть выпуклый компакт. Для всякого $x = (u, v) \in Q$ будем полагать, что $||x|| = \sqrt{||u||_1^2 + ||v||_2^2}$, где $||\cdot||_1$ и $||\cdot||_2$ — нормы в пространствах $\mathbb{R}^{n_1}$ и $\mathbb{R}^{n_2}$). Условимся обозначать $x = (u_x, v_x)$, $y = (u_y, v_y) \in Q$.

Хорошо известно, что для достаточно гладкой функции f по u и v задача (3.46) сводится к вариационному неравенству с оператором

$$G(x) = \begin{pmatrix} f'_u(u_x, v_x) \\ -f'_v(u_x, v_x) \end{pmatrix}.$$

Предложим некоторую вариацию (δ, Δ, L)-модели для вариационных неравенств, но уже на более узком классе седловых задач.

Определение 3.4.14. Будем говорить, что для некоторой постоянной $\delta > 0$ функция $\psi_\delta(x, y)$ ($\psi : \mathbb{R}^{n_1+n_2} \times \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}$) есть (δ, Δ, L)-**модель** для седловой задачи (3.46) на множестве Q, если для некоторого функционала ψ_δ при произвольных $x, y, z \in Q$ выполнены предположения (i)–(iv) определения 3.4.12, а также справедливо неравенство:

$$f(u_y, v_x) - f(u_x, v_y) \leqslant -\psi_\delta(x, y) + \delta \quad \forall x, y \in Q.$$

Из теоремы 3.4.13 вытекает

Теорема 3.4.15. *Если для седловой задачи* (3.46) *существует* (δ, Δ, L)*-модель* $\psi_\delta(x, y)$ *на множестве* Q*, то после остановки алгоритма 9 получим выходную точку*

$$\widetilde{y} = (u_{\widetilde{y}}, v_{\widetilde{y}}) := (\widetilde{u}, \widetilde{v}) := \frac{1}{S_N} \sum_{k=0}^{N_1} \frac{y^{k+1}}{L_{k+1}},$$

для которой верна оценка величины-качества решения седловой задачи:

$$\max_{v \in Q_2} f(\widetilde{u}, v) - \min_{u \in Q_1} f(u, \widetilde{v}) \leqslant \varepsilon + 2\widetilde{\delta} + \delta + \frac{1}{S_N} \sum_{k=0}^{N-1} \frac{\delta_{k+1} + \Delta_{k+1} \left\| y^{k+1} - x^{k+1} \right\|}{L_{k+1}}.$$

Заключительные замечания к главе 3

В настоящей главе рассмотрены некоторые вариации концепции неточной модели целевой функции в оптимизации, связанные с общим неравенством (3.4). Предложенный подход позволяет раздельно учитывать как погрешности целевого функционала, так и погрешность задания градиента. Предложены методы с адаптивным выбором шага, а также адаптивной настройкой величины в оценке скорости сходимости, которая определяется упомянутыми погрешностями. Преимущества использования адаптивных методов проиллюстрированы, в частности, некоторыми численными экспериментами.

Получены оценки скорости сходимости для адаптивного неускоренного градиентного метода с адаптивной настройкой на уровень гладкости задачи, а также величины погрешностей. Рассмотрен также вариант ускоренного градиентного метода для соответствующей концепции

неточной модели оптимизируемого функционала. Доказано, что для неускоренного метода не накапливаются все типы погрешностей модели. Для ускоренного метода обоснована возможность уменьшения влияния одного из параметров неточной оптимизационной модели Δ на оценку качества решения до любой приемлемой величины при накоплении величин, соответствующих постоянным значениям величин погрешностей используемой концепции модели оптимизируемого функционала. При этом показано, что адаптивность метода может на практике улучшать качество найденного решения по сравнению с полученными теоретическими оценками. Однако в полученных оценках качества решения реализована адаптивная настройка не всех параметров неточной модели. Полная адаптивная настройка на величины погрешностей возможна для искусственных неточностей, связанных с рассмотрением негладких задач. Заметим, что в разделах 3.1–3.3 мы опускаем вопрос влияния на итоговые оценки погрешности $\widetilde{\delta}$ решения вспомогательных задач на итерациях методов [22] (см. также алгоритм 1 для вариационных неравенств из предыдущей главы). Это связано с тем, что в отличие от погрешности оракула (модели) δ, по-видимому, не удаётся реализовать адаптивную настройку на величину погрешностей для вспомогательных задач $\widetilde{\delta}$. Для фиксированных величин таких погрешностей оценки скорости сходимости градиентных методов получены в [22]. В разделах 3.1–3.3 мы фокусируемся на новой части в том плане, что предложены способы адаптивной настройки не только на константы (уровень) гладкости L, а и на величины погрешностей модели δ и Δ. По-видимому, разработанные методы можно применять и при наличии погрешностей $\widetilde{\delta}$, но они не будут накапливаться в оценках как для неускоренного, так и для ускоренного методов аналогично [22]. Основные результаты данной главы опубликованы в [52, 175].

Отметим также, что обоснована применимость неускоренных процедур для относительно гладких целевых функционалов. В таком случае полученную оценку скорости сходимости $O(\varepsilon^{-1})$ можно считать оптимальной даже при отсутствии погрешностей [92]. Показано, как можно ввести аналогичное понятие неточной модели для вариационных неравенств и седловых задач и предложить аналог экстраградиентного метода с адаптивной настройкой на δ и Δ.

ГЛАВА 4

О некоторых адаптивных алгоритмических методах для задач оптимизации с близкой к линейной скоростью сходимости

Введение

В предыдущих двух разделах работы предложены методы для задач минимизации функционалов, а также для вариационных неравенств с адаптивной настройкой как на величину константы гладкости, так и на величины возможных погрешностей. Эти результаты, в основном, основаны на обобщениях концепции (δ, L)-оракула в оптимизации и её вариантах для вариационных неравенств и седловых задач, которые предложены в главе 2. Несмотря на полученные оценки с адаптивной настройкой как на уровень гладкости задачи, так и на величины погрешностей. а также неплохие результаты экспериментов для некоторых рассмотренных примеров задач, теоретически полученные оценки скорости сходимости гарантируют лишь сублинейную скорость (и хорошо известно, что эти оценки не улучшаемы). В этой связи естественно возникает проблема описания класса задач, для которых возможно разработать методы с гарантией линейной скорости сходимости (возможно, с точностью до величин, определяемых погрешностями данных). В этом плане естественно добавить к рассмотренным выше подходам к понятиям неточных оптимизационных моделей какой-то вариант условия сильной выпуклости целевого функционала или сильной монотонности оператора вариационного неравенства. В настоящей главе будут рассмотрены алгоритмические методы с адаптивной настройкой констант гладкости и соответствующих погрешностям величин, для которых можно обосновать линейную скорость сходимости.

Один из центральных рассматриваемых вопросов — описание влияния погрешностей задания (суб)градиента на оценки скорости сходимости (качества найденного решения) разрабатываемых численных

методов. Ситуацию погрешности задания градиента, в частности, описывают неравенства (3.4) и (3.3). Эти неравенства вполне аналогичны неравенству из концепции неточного оракула (3.1), однако величины $\Delta\|y-x\|$ и $\gamma\|y-x\|$ уже зависят от выбора x и y. Заменить их в (3.4) обе на постоянные величины, вообще говоря, возможно только в случае ограниченного допустимого множества задачи Q. Более того, хорошо известно, что при использовании $\widetilde{\nabla} f(x)$ из (3.2) метод может расходиться ([91], Sect. 4). Поэтому важно выделить класс задач, для которых можно получать приемлемые оценки скорости сходимости, в том числе и на неограниченных множествах. Это, в частности, мотивировало часть исследований настоящей главы. Напомним, для сильно выпуклого целевого функционала с липшицевым градиентом известно, что градиентный метод сходится с линейной скоростью. Весьма интересен вопрос о том, насколько можно условие сильной выпуклости ослабить. В этом плане весьма известен подход, основанный на использовании вместо сильной выпуклости условия градиентного доминирования Поляка–Лоясиевича [45] (см. также недавнюю работу [114] и имеющиеся там ссылки)

$$f(x) - f(x_*) \leqslant \frac{1}{2\mu}\|\nabla f(x)\|_2^2, \tag{4.1}$$

где $f(x_*) = f^*$ — искомое оптимальное значение f, а $\|\cdot\|_2$ — евклидова норма. Известно, что данное неравенство в предположении липшицевости градиента позволяет получить оценку скорости сходимости

$$f(x^N)-f(x_*) \leqslant \left(1-\frac{\mu}{L}\right)^N \left(f(x^0)-f(x_*)\right) \leqslant \exp\left(-\frac{\mu}{L}N\right)\left(f(x^0)-f(x_*)\right). \tag{4.2}$$

При этом мы рассматриваем некоторое ослабление условия липшицевости градиента

$$f(y) \leqslant f_\delta(x) + \langle \widetilde{\nabla} f(x), y-x\rangle + \frac{L}{2}\|y-x\|_2^2 + \Delta\|y-x\|_2 + \delta \quad \forall\, x, y \in Q$$

для некоторых $\delta > 0$ и $\Delta > 0$:

$$f_\delta(x) \in [f(x)-\delta; f(x)] \ \forall x \in Q.$$

Например, это предположение естественно в случае, если значения f несколько отличаются от значений некоторой достаточно гладкой функции $\widetilde{f}$, удовлетворяющей условию Липшица градиента (при этом

$\widetilde{\nabla} f(x)$ — некоторое возмущенное с точностью Δ значение градиента $\nabla f(x)$). По сути, в этой части работы левая часть неравенства (3.4) заменяется условием градиентного доминирования. Мы предлагаем метод с адаптивным подбором шага с настройкой на величины L, Δ и δ и показываем оценку скорости сходимости, аналогичную (4.2). В частности, запуск предлагаемого метода (алгоритм 10) не предполагает знания никакой верхней оценки L и может применяться для задач с неточным заданием градиента на неограниченных допустимых множествах. Итак, в разделе 4.1 предложен адаптивный градиентный метод для целевых функционалов с липшицевым градиентом (а также некоторой релаксацией этого условия), удовлетворяющих условию Поляка–Лоясиевича. При этом учитывается возможность неточного задания градиента и предлагается адаптивная настройка работы метода на основные входные параметры. Обоснована линейная сходимость метода с точностью до величины, связанной с погрешностью. На примере задачи решения матричного уравнения экспериментально проведено сравнение скорости сходимости адаптивного и его неадаптивного аналога при наличии погрешности задания градиента. Также отметим, что использование данного подхода даёт возможность оценивать качество решения по значению целевой функции в начальной точке $f(x^0) \geqslant f(x^0) - f(x_*)$ без знания оценки расстояния от x^0 до точного решения, если функция неотрицательна ($f(x) \geqslant 0$).

Подход первого раздела главы позволяет говорить о применимости методики для некоторых невыпуклых задач, но при этом он существенно основан на выборе евклидовой нормы и не проработан в модельной общности (и пока что даже для композитной оптимизации). В двух последующих разделах речь пойдёт об аналогах (δ, L, μ)-оракула О. Деволдера–Ф. Глинера–Ю. Е. Нестерова, предполагающих как рассмотрение модельной общности, так и условий относительной гладкости [67] и относительной сильной выпуклости [124]. Заметим, что использование (δ, L, μ)-оракула может позволить работать с задачами при аддитивной погрешности задания (суб)градиента (Δ) на не обязательно ограниченных допустимых множествах (более детально это обсуждается в [91]), но при этом выражение для параметра δ (как и оценки скорости сходимости) содержит величину вида $\frac{\Delta^2}{\mu}$, которая может быть довольно большой при достаточно малых μ. Также подход [91] существенно основан на сильной выпуклости прокс-функции, что невозмож-

но гарантировать для относительно гладких и/или сильно выпуклых оптимизационных задач. В разделах 4.2 и 4.3 предложены альтернативные подходы, которые могут позволить обходить такие проблемы с помощью двух параметров неточностей (δ и Δ), адаптивной настройки этих параметров в ходе работы методов, а также расширить класс задач (модельная общность, относительно гладкие и относительно сильно выпуклые задачи).

Так, в разделе 4.2 введен аналог (δ, L, μ)-оракула — (δ, L, μ)-модель целевой функции и предложен адаптивный градиентный метод с гарантией близкой к линейной (с точностью до параметров, соответствующих погрешностям) скорости сходимости. Следующий раздел 4.3 посвящён уже рассмотренному в предыдущей главе аналогу концепции (δ, Δ, L)-модели целевого функционала, а именно — понятию (δ, Δ, L, μ)-модели и методу градиентного типа с адаптацией не только к уровню гладкости, но и к параметрам погрешностей. Обоснована близкая к линейной скорость сходимости метода с точностью до величин, соответствующих погрешностям.

В заключительных двух разделах главы рассмотрены методы для специального типа задач выпуклого программирования с одним или двумя функционалами ограничений. Небольшое количество ограничений приводит к малой размерности двойственной задачи, ввиду чего для близкой к линейной скорости сходимости вполне достаточно потребовать сильную выпуклость лишь для целевого функционала. На базе комбинации метода дихотомии для одномерной двойственной задачи и методов градиентного типа для вспомогательных многомерных задач предложен подход к задачам выпуклого программирования с одним или двумя функционалами ограничений. Метод основан на введённом адаптивном критерии остановки по зазору двойственности. Однако в качестве особенности предлагаемого подхода стоит подчеркнуть, что мы рассматриваем весьма общую постановку, которая не предполагает явного задания двойственной задачи. Как следствие, в полученных итоговых оценках скорости сходимости учтена неточность решения вспомогательных задач. Для задач с двумя функционалами ограничений двойственная задача уже имеет размерность 2 и для неё предлагается подход, основанный на некотором аналоге дихотомии для минимизации выпуклой функции двух переменных на квадрате. Этот метод ранее предложен Ю. Е. Нестеровым в предположении точного решения

вспомогательных подзадач. В работе [44] он уже детально теоретически исследован с учётом возможных погрешностей при решении вспомогательных подзадач, погрешностей задания градиента целевого функционала (что важно для приложений выделенному классу задач выпуклого программирования, если нет возможности явно построить двойственную задачу). На базе полученной в [44] оценки скорости сходимости для аналога метода дихотомии на квадрате с учётом погрешности нахождении градиента предложен подход к задаче выпуклого программирования с двумя функциональными ограничениями. Показана линейная скорость сходимости предложенной методики для достаточно гладких задач с сильно выпуклым целевым функционалом как в случае одного, так и двух функционалов ограничений.

4.1 Адаптивный метод для минимизации функций, удовлетворяющих условию градиентного доминирования при неточном задании целевой функции и градиента

Теперь предложим подход к задаче минимизации, вообще говоря, невыпуклых функций с неточно заданным градиентом. При этом метод предполагает адаптивную настройку на некоторые параметры, в том числе связанные с величину погрешности задания градиента. Пусть рассматривается задача минимизации функции на всем пространстве $f : \mathbb{R}^n \to \mathbb{R}$ с евклидовой нормой $\|\cdot\| = \|\cdot\|_2$. При этом:

(i) существует $x_* \in \mathbb{R}^n$:

$$f(x_*) = \min_{x \in \mathbb{R}^n} f(x) =: f^*; \tag{4.3}$$

(ii) выполнено условие Поляка–Лоясиевича или (PL)-условие:

$$f(x) - f^* \leqslant \frac{1}{2\mu} \|\nabla f(x)\|_2^2 \quad \forall x \in \mathbb{R}^n; \tag{4.4}$$

(iii) неравенство

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L \|y - x\|_2^2}{2} \quad \forall x, y \in \mathbb{R}^n.$$

Если предположить, что в каждой точке $x \in \mathbb{R}^n$ доступно приближенное значение $\widetilde{\nabla} f(x)$ градиента $\nabla f(x)$: $\left\|\widetilde{\nabla} f(x) - \nabla f(x)\right\|_2 \leqslant \Delta \ \ \forall x \in \mathbb{R}^n$ при некотором фиксированном $\Delta > 0$, то верно

$$f(y) \leqslant f(x) + \left\langle \widetilde{\nabla} f(x), y - x \right\rangle + \frac{L}{2}\|y - x\|_2^2 + \Delta\|y - x\|_2 \quad \forall x, y \in \mathbb{R}^n.$$

Далее для удобства будем обозначать $g_x := \|\nabla f(x)\|_2$ и $\widetilde{g}_x := \left\|\widetilde{\nabla} f(x)\right\|_2$.

К задаче (4.3) будем применять градиентный метод вида

$$x^{k+1} = x^k - h_k \widetilde{\nabla} f(x^k), \tag{4.5}$$

$k = 0, 1, 2, \ldots$ и $h_k > 0$. При этом h_k выберем так, чтобы

$$f(x^{k+1}) \leqslant f(x^k) + \left\langle \widetilde{\nabla} f(x^k), x^{k+1} - x^k \right\rangle + \frac{L\left\|x^{k+1} - x^k\right\|_2^2}{2} + \\ + \Delta_{k+1}\left\|x^{k+1} - x^k\right\|_2, \tag{4.6}$$

где $\Delta_{k+1} > 0$ — адаптивно подбираемая величина. В начале каждой итерации $\Delta_{k+1} := \frac{\Delta_k}{2}$, а далее Δ_{k+1} $(k = 0, 1, 2, \ldots)$ увеличивается в два раза и процедура (4.5) повторяется до тех пор, пока не выполняется (4.6).

Ясно, что (4.6) заведомо верно при $\Delta_{k+1} \geqslant \Delta$. Поэтому аналогично пункту 2) доказательства теоремы 3.1.4 проверяется, что за конечное число таких шагов (4.6) будет выполнено на любой итерации $(k = 0, 1, 2, \ldots)$, после чего $f(x^{k+1}) - f(x^k) \leqslant \varphi(h_k)$, где $\varphi(h) = -h\widetilde{g}_{x^k}^2 + \frac{Lh^2}{2}\widetilde{g}_{x^k}^2 + \Delta_{k+1}\widetilde{g}_{x^k}$. Выберем шаг h_k так, чтобы минимизировать величину $\varphi(h_k)$, то есть $\varphi'(h_k) = 0$ и

$$h_k = \frac{1}{L} - \frac{\Delta_{k+1}}{L\widetilde{g}_{x^k}}. \tag{4.7}$$

Далее по умолчанию считаем, что $h_k > 0$. Если это не так, то $\widetilde{g}_{x^k} \leqslant \Delta_{k+1} \leqslant 2\Delta$ и мы получаем, что в текущей точке целевая функция имеет достаточно малый градиент $\widetilde{g}_{x^k} \leqslant 3\Delta$, что гарантированно приводит к приемлемому для достаточно малых Δ качеству решения по функции

$$f(x^k) - f(x_*) \leqslant \frac{9\Delta^2}{2\mu}. \tag{4.8}$$

Если же величина шага h_k из (4.7) положительна, то

$$\varphi(h_k) =$$

$$= -\left(\frac{1}{L} - \frac{\Delta_{k+1}}{\widetilde{g}_{x^k} L}\right)\widetilde{g}_{x^k}^2 + \frac{L\widetilde{g}_{x^k}^2}{2}\left(\frac{1}{L} - \frac{\Delta_{k+1}}{\widetilde{g}_{x^k} L}\right)^2 + \left(\frac{1}{L} - \frac{\Delta_{k+1}}{\widetilde{g}_{x^k} L}\right)\widetilde{g}_{x^k}\Delta_{k+1} =$$

$$= -\frac{\widetilde{g}_{x^k}^2}{L} + \frac{\Delta_{k+1}\widetilde{g}_{x^k}}{L} + \frac{\widetilde{g}_{x^k}^2}{2L} - \frac{\Delta_{k+1}\widetilde{g}_{x^k}}{L} + \frac{\Delta_{k+1}^2}{2L} + \frac{\widetilde{g}_{x^k}\Delta_{k+1}}{L} - \frac{\Delta_{k+1}^2}{L} =$$

$$= -\frac{\widetilde{g}_{x^k}^2}{2L} + \frac{\widetilde{g}_{x^k}\Delta_{k+1}}{L} - \frac{\Delta_{k+1}^2}{2L} = -\frac{1}{2L}\left(\widetilde{g}_{x^k} - \Delta_{k+1}\right)^2.$$

Поэтому (4.6) означает, что

$$f(x^{k+1}) - f(x^k) \leqslant -\frac{1}{2L}\left(\widetilde{g}_{x^k} - \Delta_{k+1}\right)^2 \leqslant -\frac{1}{2L}\left(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}\right) g_{x^k}^2, \quad (4.9)$$

поскольку

$$|\widetilde{g}_{x^k} - g_{x^k}| \leqslant \left\|\widetilde{\nabla} f(x^k) - \nabla f(x^k)\right\|_2 \leqslant \Delta$$

и $\widetilde{g}_{x^k} + \Delta \geqslant g_{x^k}$. Неравенство (4.9) означает, что $\forall k = 0, 1, 2, \ldots$

$$f(x^k) - f(x^{k+1}) \geqslant \frac{1}{2L}\left(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}\right)^2 g_{x^k}^2 \geqslant \frac{\mu}{L}\left(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}\right)^2 \left(f(x^k) - f^*\right)$$

ввиду (PL)-условия (4.4). Поэтому

$$f(x^{k+1}) - f(x_*) \leqslant \left(1 - \frac{\mu}{L}\left(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}\right)^2\right)\left(f(x^k) - f(x_*)\right),$$

откуда

$$f(x^{k+1}) - f^* \leqslant \prod_{i=0}^{k}\left(1 - \frac{\mu}{L}\left(\frac{\widetilde{g}_{x^i} - \Delta_{i+1}}{\widetilde{g}_{x^i} + \Delta}\right)^2\right)\left(f(x^0) - f^*\right). \quad (4.10)$$

Можно считать $\mu \leqslant L$ и ввиду $\widetilde{g}_{x^i} - \Delta_{i+1} < \widetilde{g}_{x^i} + \Delta$ в (4.10) справа входит произведение $k+1$ числа, каждое из которых меньше 1. Адаптивность подбора $\Delta_{k+1} \leqslant 2\Delta$ на каждой итерации может привести к увеличению дроби $\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}$ и уменьшению множителей в (4.10), что потенциально улучшает оценку (такого типа ситуация проиллюстрирована численными экспериментами далее) по сравнению с неадаптивным вариантом

$$f(x^{k+1}) - f^* \leqslant \prod_{i=0}^{k}\left(1 - \frac{\mu}{L}\left(\frac{\widetilde{g}_{x^i} - \Delta}{\widetilde{g}_{x^i} + \Delta}\right)^2\right)\left(f(x^0) - f^*\right). \quad (4.11)$$

Замечание 4.1.1. Покажем, как из (4.10) можно вывести более конкретную оценку скорости сходимости. Если известно Δ, то без уменьшения общности рассуждений можно положить $\Delta_{k+1} := \min\{\Delta_{k+1}, \Delta\}$. Ясно, что $\forall k = 0, 1, 2, \ldots$

$$\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta} \geqslant \frac{\widetilde{g}_{x^k} - \Delta}{\widetilde{g}_{x^k} + \Delta} = 1 - \frac{2\Delta}{\widetilde{g}_{x^k} + \Delta}.$$

Пусть $\widetilde{g}_{x^k} \geqslant C\Delta$ для некоторой постоянной $C > 1$. Тогда

$$1 - \frac{2\Delta}{\widetilde{g}_{x^k} + \Delta} \geqslant 1 - \frac{2}{C+1} = \frac{C-1}{C+1} > 0$$

и (4.10) принимает вид

$$f(x^{k+1}) - f^* \leqslant \left(1 - \frac{\mu}{L}\left(\frac{C-1}{C+1}\right)^2\right)^{k+1} \left(f(x^0) - f^*\right), \tag{4.12}$$

что означает сходимость со скоростью геометрической прогрессии. Если же для некоторого k верно $\widetilde{g}_{x^k} < C\Delta$, то $g_{x^k} < C\Delta + \Delta = \Delta(C+1)$ и согласно (PL)-условию (4.4)

$$f(x^k) - f^* < \frac{(C+1)^2\Delta^2}{2\mu}. \tag{4.13}$$

Таким образом, доказано что либо невязка $\min\limits_k f(x^k) - f^*$ убывает при увеличении k со скоростью геометрической прогрессии (см. (4.14)), либо (см. (4.13)) эта невязка ограничена сопоставимой с Δ^2 величиной (как показывает (4.8) это верно и если $h_k \leqslant 0$). Аналогичные вычислительные гарантии можно обосновать и для следующего метода с адаптивным выбором не только параметров погрешности на итерациях метода, но и соответствующих L величин.

Для указанного алгоритма будет верна такая оценка, обоснование которой аналогично (4.11)

$$f(x^{k+1}) - f^* \leqslant \prod_{i=0}^{k} \left(1 - \frac{\mu}{L_{i+1}}\left(\frac{\widetilde{g}_{x^i} - \Delta_{k+1}}{\widetilde{g}_{x^i} + \Delta}\right)^2\right) \left(f(x^0) - f^*\right).$$

Описанная в замечании 4.1.1 схема рассуждений приводит нас к следующему результату.

Алгоритм 10 Адаптивный градиентный метод для функций, удовлетворяющих (PL)-условию.

Require: x^0 — начальная точка, параметры $\Delta_0,\ L_0$ $(2\mu \leqslant L_0 < 2L,\ \Delta_0 \leqslant 2\Delta)$.

1: $L_{k+1} := \max\{\mu, {}^{L_k}/_2\},\ \Delta_{k+1} := {}^{\Delta_k}/_2$.

2: $x^{k+1} = x^k - h_k \widetilde{\nabla} f(x^k)$,

$$h_k = \frac{1}{L_{k+1}} - \frac{\Delta_{k+1}}{L_{k+1}\widetilde{g}_{x^k}},\ \widetilde{g}_{x^k} = \left\|\widetilde{\nabla} f(x^k)\right\|_2.$$

3: **repeat**

4: **if** $f(x^{k+1}) \leqslant f(x^k) + \left\langle \widetilde{\nabla} f(x^k), x^{k+1} - x^k \right\rangle + \frac{L_{k+1}}{2}\left\|x^{k+1} - x^k\right\|_2^2 + \Delta_{k+1}\left\|x^{k+1} - x^k\right\|_2$ **then**

5: $k := k+1$ и выполнение п. 1.

6: **else**

7: $L_{k+1} := 2 \cdot L_{k+1};\ \Delta_{k+1} := 2 \cdot \Delta_{k+1}$ и выполнение п. 2.

8: **end if**

9: **until** $k \geqslant N$

Ensure: x^{k+1}.

Теорема 4.1.2. *Пусть для некоторого натурального k $h_i > 0$ при $i \leqslant k$ и верно $\Delta_{k+1} \leqslant \Delta$. Тогда после k итераций алгоритма 10 для всякого $C > 1$ будет выполняться одно из двух неравенств*

$$f(x^{k+1}) - f(x_*) \leqslant \left(1 - \frac{\mu}{\widehat{L}}\left(\frac{C-1}{C+1}\right)^2\right)^{k+1}(f(x^0) - f(x_*)) \tag{4.14}$$

где $\widehat{L} \leqslant 2L\max\left\{1, \frac{2\Delta}{\Delta_0}\right\}$ или

$$\min_{i=\overline{1,k+1}} f(x^i) - f^* < \frac{(C+1)^2\Delta^2}{2\mu}.$$

Доказательство. Отметим лишь, что неравенство (4.14) следует из того, что на каждой итерации алгоритма 10 в ходе проверок выполнения критерия выхода из итерации (пункт 4 листинга алгоритма 10) потенциально возможно увеличение L_{k+1} не более чем до $\widehat{L}$ в случае, если ввиду малости Δ_{k+1} критерий выхода из итерации не выполнен. □

Замечание 4.1.3. Если на снять ограничение п. 1 листинга алгоритма

10 и какой-то итерации $L_{k+1} < \mu$, то возможно

$$1 - \frac{\mu}{L_{k+1}} \left(\frac{\widetilde{g}_{x^k} - \Delta}{\widetilde{g}_{x^k} + \Delta} \right) < 0.$$

В этом случае

$$f(x^{k+1}) - f^* \leqslant \left(1 - \frac{\mu}{L_{k+1}} \left(\frac{\widetilde{g}_{x^i} - \Delta}{\widetilde{g}_{x^i} + \Delta} \right)^2 \right) \left(f(x^k) - f^* \right)$$

и тогда $f(x^{k+1}) = f(x^k) = f^*$, что гарантирует достижение точного решения.

Замечание 4.1.4. Можно рассматривать вместо (4.5) более слабое условие

$$f(y) \leqslant f_\delta(x) + \langle \widetilde{\nabla} f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2 + \Delta \|y - x\|_2 + \delta \quad \forall\, x, y \in Q \tag{4.15}$$

для некоторого $\delta > 0$, а также $f_\delta(x) \in [f(x) - \delta, f(x)] \ \forall x \in Q$. Например, это актуально в случае, если значения f несколько отличаются от значений некоторой достаточно гладкой функции $\widetilde{f}$, удовлетворяющей (4.5) (при этом $\widetilde{\nabla} f(x)$ — некоторое возмущенное с точностью Δ значение градиента $\nabla \widetilde{f}(x)$).

В таком случае рассмотрим алгоритм 10 с видоизмененным критерием выхода из итерации (пункт 4 листинга этого алгоритма)

$$\begin{gathered} f_\delta(x^{k+1}) \leqslant f_\delta(x^k) + \langle \widetilde{\nabla} f(x^k), x^{k+1} - x^k \rangle + \\ \frac{L_{k+1} \|x^{k+1} - x^k\|_2^2}{2} + \Delta_{k+1} \|x^{k+1} - x^k\|_2 + \delta_{k+1}. \end{gathered} \tag{4.16}$$

При этом предполагается адаптивный подбор величин Δ_{k+1} и δ_{k+1} при заданных изначально $L_0 \leqslant 2L$, $\Delta_0 \leqslant \Delta$ и $\delta_0 \leqslant 2\delta$. Тогда на каждой итерации вместо неравенства (4.9) будет верно

$$f(x^k) - f(x^{k+1}) + \delta + \delta_{k+1} \geqslant \frac{\mu}{L_{k+1}} \left(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta} \right)^2 \left(f(x^k) - f^* \right),$$

откуда аналогично (4.10) имеем

$$f(x^{k+1}) - f^* \leqslant \left(1 - \frac{\mu}{L_{k+1}} \left(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta} \right)^2 \right) \left(f(x^k) - f^* \right) + \delta + \delta_{k+1} \leqslant$$

$$\leqslant \left(1 - \frac{\mu}{L_{k+1}}\left(\frac{\widetilde{g}_{x^k} - \Delta_{k+1}}{\widetilde{g}_{x^k} + \Delta}\right)^2\right)\left(1 - \frac{\mu}{L_k}\left(\frac{\widetilde{g}_{x^{k-1}} - \Delta_k}{\widetilde{g}_{x^{k-1}} + \Delta}\right)^2\right)\cdot$$

$$\cdot \left(f(x^{k-1}) - f^*\right) + \delta + \delta_k\left(1 - \frac{\mu}{L_{k+1}}\left(\frac{\widetilde{g}_{x^{k-1}} - \Delta_k}{\widetilde{g}_{x^{k-1}} + \Delta}\right)^2\right) + \delta + \delta_{k+1} \leqslant \ldots \leqslant$$

$$\leqslant \prod_{i=0}^{k}\left(1 - \frac{\mu}{L_{i+1}}\left(\frac{\widetilde{g}_{x^i} - \Delta_{i+1}}{\widetilde{g}_{x^i} + \Delta}\right)^2\right)\left(f(x^0) - f^*\right) +$$

$$+ \sum_{i=0}^{k-1}(\delta + \delta_{i+1})\prod_{j=i}^{k}\left(1 - \frac{\mu}{L_{j+1}}\left(\frac{\widetilde{g}_{x^j} - \Delta_{j+1}}{\widetilde{g}_{x^j} + \Delta}\right)^2\right) + \delta + \delta_{k+1}.$$

Полученная оценка выглядит несколько громоздко. Конкретизируем её при постоянном $L_{i+1} = L$, $\Delta = 0$ и $\delta_{i+1} \leqslant \delta$ $(i \geqslant 0)$:

$$f(x^{k+1}) - f^* \leqslant \left(1 - \frac{\mu}{L}\right)^{k+1}\left(f(x^0) - f^*\right) + \sum_{i=0}^{k}(\delta + \delta_{i+1})\left(1 - \frac{\mu}{L}\right)^{k-i} \leqslant$$

$$\leqslant \left(1 - \frac{\mu}{L}\right)^{k+1}\left(f(x^0) - f^*\right) + 2\delta\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^{k-i} =$$

$$= \left(1 - \frac{\mu}{L}\right)^{k+1}\left(f(x^0) - f^*\right) + \frac{3\delta L}{\mu}.$$

Данное неравенство приводит к таким выводам. С одной стороны мы видим, что величина, связанная с погрешностью δ, ограничена. Однако, она может быть довольно немалой при большом значении числа обусловленности $\frac{L}{\mu}$. Это показывает также, что замена слагаемого $\Delta\|y - x\|_2$ в (4.15) на $\frac{\Delta^2 + \|y-x\|_2^2}{2}$ может привести к ухудшению оценки качества решения при достаточно большом $\frac{L}{\mu}$.

Замечание 4.1.5. Аналогично второй части доказательства теоремы 3.1.4 можно проверить, что трудоемкость итерации адаптивного алгоритма 10 сопоставима с трудоемкостью аналогичного неадаптивного метода.

Замечание 4.1.6. Важно отметить, что оценки скорости сходимости выше обоснованы лишь при условии положительных величин шагов $h_k > 0$, что не всегда верно при $\Delta_{k+1} > 0$. Чтобы избежать этой проблемы, вполне возможно применять адаптивный алгоритм 10 с видоизмененным критерием выхода из итерации (4.16) при условии $\Delta_{k+1} = 0$

($k = 0, 1, 2, ...$). Однако при этом Δ может быть положительным. Действительно, если предположить, что в каждой точке x доступно приближенное значение целевого функционала $f_\delta(x)$:

$$f(x) - \delta \leqslant f_\delta(x) \leqslant f(x),$$

а также Δ-приближенное значение $\widetilde{\nabla} f(x)$ (суб)градиента $\nabla f(x)$:

$$\left\| \widetilde{\nabla} f(x) - \nabla f(x) \right\|_2 \leqslant \Delta \ \ (\Delta > 0)$$

и верно неравенство для некоторого $L > 0$

$$f_\delta(y) \leqslant f(y) \leqslant f_\delta(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2 + \delta \ \ \forall x, y \in Q,$$

то по неравенствам Коши-Буняковского и Евклида

$$\left| \left\langle \nabla f(x) - \widetilde{\nabla} f(x), y - x \right\rangle \right| \leqslant \left\| \nabla f(x) - \widetilde{\nabla} f(x) \right\|_2 \|y - x\|_2 \leqslant$$

$$\leqslant \frac{1}{2L} \left\| \nabla f(x) - \widetilde{\nabla} f(x) \right\|_2^2 + \frac{L}{2} \|y - x\|_2^2 \leqslant \frac{\Delta^2}{2L} + \frac{L}{2} \|y - x\|_2^2 \ \ \forall x, y \in Q.$$

Поэтому имеем $\forall x, y \in Q$

$$f_\delta(y) \leqslant f(y) \leqslant f_\delta(x) + \left\langle \widetilde{\nabla} f(x), y - x \right\rangle + L \|y - x\|_2^2 + \delta + \frac{\Delta^2}{2L}.$$

Поэтому адаптивный метод при $\Delta_{k+1} = 0$ ($\forall k \geqslant 0$) и доказанная для него оценка скорости сходимости применимы к ситуации неточного задания градиента. Однако при этом в теоретической оценке скорости сходимости может появиться дополнительное слагаемое вида $O\left(\frac{\Delta^2}{2\mu}\right)$, которое потенциально велико при малых значениях μ.

В заключение данного раздела рассмотрим пример численных экспериментов, демонстрирующих преимущество метода с адаптивной настройкой параметров по сравнению с неадаптивным вариантом.

Пример 4.1.7. Пусть дана матрица A размера 1000×1000 и вектор $b \in \mathbb{R}^{1000}$. Предположим, что главная диагональ матрицы A заполнена случайными целыми числами из интервала $[1, 20]$. Также 10 случайно выбранных элементов данной матрицы — целые числа из интервала $[1, 20]$. Рассмотрим задачу решения матричного уравнения, которая

в случае разрешимости равносильна задаче минимизации выпуклого функционала $f(x) = 0.5\|Ax - b\|_2^2$ на всём пространстве $\mathbb{R}^{1000}$.

Указанная функция μ-сильно выпукла и имеет L-липшицев градиент, где μ — наименьшее положительное собственное число матрицы A^TA, L — наибольшее собственное число A^TA (A^T — матрица, транспонированная к A). Выбрана точка старта $x^0 = (0.1, \ldots, 0.1)$. Полагаем, что вместо точного значения функционала нам доступно его приближение $f_{\widetilde{\delta}}(x)$ — некоторое возмущенное значение ($f_{\widetilde{\delta}}(x) \in [f(x) - \widetilde{\delta}; f(x)]$ при всяком x) с точностью $\widetilde{\delta}$, где $\widetilde{\delta} \in (0, \delta)$ при $\delta = 0.01$). Погрешность задания градиента целевого функционала по норме ограничена величиной $\Delta = 0.25$. Сравниваются прежде всего значения функции и величины теоретических оценок в правых частях неравенств

$$f(x^{k+1}) - f^* \leqslant \prod_{i=0}^{k} \left(1 - \frac{\mu}{L}\left(\frac{\widetilde{g}_{x^i} - \Delta}{\widetilde{g}_{x^i} + \Delta}\right)^2\right) f(x^0) +$$

$$+ \sum_{i=0}^{k-1} 2\delta \prod_{j=i}^{k} \left(1 - \frac{\mu}{L}\left(\frac{\widetilde{g}_{x^j} - \Delta}{\widetilde{g}_{x^j} + \Delta}\right)^2\right) + 2\delta,$$

и

$$f(x^{k+1}) - f^* \leqslant \prod_{i=0}^{k} \left(1 - \frac{\mu}{L_{i+1}}\left(\frac{\widetilde{g}_{x^i} - \Delta_{i+1}}{\widetilde{g}_{x^i} + \Delta}\right)^2\right) f(x^0) +$$

$$+ \sum_{i=0}^{k-1} (\delta + \delta_{i+1}) \prod_{j=i}^{k} \left(1 - \frac{\mu}{L_{j+1}}\left(\frac{\widetilde{g}_{x^j} - \Delta_{j+1}}{\widetilde{g}_{x^j} + \Delta}\right)^2\right) + \delta + \delta_{k+1},$$

которые определяют качество решения для алгоритма 13 и его неадаптивного варианта. Усредненные результаты 5 экспериментов (для разных матриц A и векторов b) представлены в таблицах 4.1–4.3. В частности, в таблице 4.1 представлены результаты экспериментов для вектора b, состоящего из чисел, имеющих стандартное гамма-распределение, в таблице 4.2 — для вектора b, представленного целыми числами из интервала $[-10; 10]$, в таблице 4.3 — для вектора b, представленного числами из полуинтервала $[0; 1)$. Как видим, при одинаковом количестве итераций и сопоставимых временных затратах, адаптивный метод гарантирует существенно лучшее качество найденного решения. Отрицательность $f_{\widetilde{\delta}}(x)$ возможна как раз ввиду того, что это неточное значение f.

Ввиду адаптивности метода как по параметрам L_k, так и по параметру неточности δ, можно говорить о применимости разработанного

Таблица 4.1. Результаты сравнения работы адаптивного и неадаптивного варианта рассматриваемого метода (случайные данные генерируются с использованием стандартного гамма-распределения).

	Адаптивный		Неадаптивный	
N	$f_{\widetilde{\delta}\ \min}$	Время, с	$f_{\widetilde{\delta}(x)\ \min}$	Время, с
1000	-0.00434	109	0.67651	33
2000	-0.00841	212	0.05211	67
3000	-0.00944	320	0.0272	99
4000	-0.00978	420	0.02291	131
5000	-0.00986	525	0.02166	161
6000	-0.00988	628	0.02119	188
7000	-0.00988	725	0.02099	215
8000	-0.00989	825	0.0209	242
9000	-0.0099	927	0.02083	269
10000	-0.0099 (оценка: 4.0)	1028	0.02081 (оценка: 236.1)	295

Таблица 4.2. Результаты сравнения работы адаптивного и неадаптивного варианта рассматриваемого метода (случайные данные генерируются с использованием целых чисел из интервала $[-10; 10]$).

	Адаптивный		Неадаптивный	
N	$f_{\widetilde{\delta}\ \min}$	Время, с	$f_{\widetilde{\delta}(x)\ \min}$	Время, с
1000	0.25388	213	14.78208	68
2000	0.02253	422	2.2755	134
3000	-0.00317	632	1.225	198
4000	-0.00811	829	0.86423	262
5000	0.00067	1028	0.69942	322
6000	-0.00954	1220	0.61284	378
7000	-0.00961	1427	0.56266	433
8000	-0.00962	1624	0.53163	487
9000	-0.00963	1821	0.51131	540
10000	-0.00963 (теор. оценка: 186.0)	1028	0.49764 (теор. оценка: 4422.3)	295

Таблица 4.3. Результаты сравнения работы адаптивного и неадаптивного варианта рассматриваемого метода (случайные данные генерируются с использованием чисел из полуинтервала $[0; 1)$).

	Адаптивный		Неадаптивный	
N	$f_{\widetilde{\delta}\ \min}$	Время, с	$f_{\widetilde{\delta}(x)\ \min}$	Время, с
1000	0.01211	214	0.50635	69
2000	-0.00298	423	0.21837	134
3000	-0.00715	632	0.18413	198
4000	-0.00877	830	0.17643	260
5000	-0.00935	1030	0.17433	315
6000	-0.00958	1224	0.17352	369
7000	-0.00968	1432	0.17323	422
8000	-0.00972	1631	0.17293	473
9000	-0.00972	1830	0.17277	523
10000	-0.0097 (теор. оценка: 2.6)	1028	0.1727 (теор. оценка: 50.0)	295

подхода к задачам негладкой оптимизации (правда без гарантированного достижения оптимальных теоретических оценок сложности, что связано с наличием в теоретической оценке числа обусловленности при параметре неточности δ). Тем не менее, в силу адаптивного подбора (локальных) параметров гладкости L_k на итерациях метода по выведенной теоретической оценке скорости сходимости вполне возможно делать выводы о качестве найденного решения. Важно отметить, что использование предлагаемого в настоящем пункте подхода и полученных теоретических оценок даёт возможность оценивать качество найденного решения по значению неотрицательной в $\mathbb{R}^n$ (если есть ограниченность целевой функции снизу, то возможно модифицировать это условие) целевой функции в начальной точке $f(x^0) \geqslant f(x^0) - f(x_*)$ без знания оценки расстояния от x^0 до точного решения задачи.

Рассмотрим некоторый пример такого типа, в котором $f_\delta = f$ и адаптивная настройка проводится только по параметрам L_k и δ_k, полагаем $\Delta_k = 0$ для всякого $k \geqslant 0$. Здесь параметр δ уже играет роль искусственной неточности, которая вносится в оптимизационную модель для негладкой задачи. Возможность адаптивной настройки L_k и δ_k гарантирует применимость полученных теоретических оценок скоро-

Таблица 4.4. Результаты сравнения работы рассмотренного алгоритма.

N	Оценка	$f_{\min}(x)$	Время, с
1000	51.26256	248.68009	229
2000	2.24426	248.4881	438
3000	0.66699	248.48333	650
4000	0.92416	248.47849	853
5000	0.61537	248.47849	1060
6000	0.51292	248.47521	1260
7000	0.71772	248.47393	1451
8000	0.71772	248.46579	1643
9000	0.61532	248.46579	1836
10000	0.51292	248.46579	2033

сти сходимости, содержащих эти величины, к негладким задачам.

Пример 4.1.8. Пусть в пространстве $\mathbb{R}^n$ (размерность $n = 1000$) задано m точек $A_k = (a_{1k}, a_{2k}, \ldots, a_{nk})$ $(k = 1, 2, \ldots, 100)$ (координаты точек A_k принимают значения -1 и 0), для которых в n-мерном евклидовом пространстве $\mathbb{R}^n$ необходимо найти такую точку $X = (x_1, x_2, \ldots, x_n)$, чтобы целевая функция

$$f(x) := \max_{k=\overline{1,m}} \left\{(x_1 - a_{1k})^2 + (x_2 - a_{2k})^2 + \ldots + (x_n - a_{nk})^2\right\}$$

принимала наименьшее значение на единичном шаре с центром в нуле. Указанная функция μ-сильно выпукла при $\mu = 2$. В то же время глобальная константа L бесконечна (поскольку целевая функция негладкая), но за счёт адаптивного подбора на итерациях параметров L_k можно наблюдать достижение приемлемого качества решения. Выбрана точка старта $x^0 = (1, \ldots, 1)$. В таблице 4.4 приведены усредненные результаты 5 экспериментов со случайным выбором координат точек для указанного количества итераций. Для оценки качества решения используются именно доказанные теоретические оценки скорости сходимости для предложенного алгоритма ($\Delta = \Delta_k = 0$ при $k = 0, 1, 2, ...$), поскольку нет никакой априорной информации о точном решении и приблизительном значении f^*.

4.2 Адаптивный градиентный спуск для задач минимизации функционалов, допускающих (δ, L, μ)-модели целевой функции в произвольной запрошенной точке.

4.2.1 Понятие (δ, L, μ)-модели целевой функции. Адаптивный градиентный метод Введем понятие (δ, L, μ)-модели целевого функционала минимизационной задачи, которая аналогична общему понятию (δ, L)-модели функции из [22]. Концепция (δ, L, μ)-модели есть обобщение (δ, L, μ)-оракула из работы [91].

Определение 4.2.1. Будем говорить, что функция f допускает (δ, L, μ)-модель в точке $x \in Q$, если для любого $y \in Q$ верно

$$\mu V(y, x) \leqslant f(y) - (f_\delta(x) + \psi(y, x)) \leqslant LV(y, x) + \delta, \tag{4.17}$$

где $\psi(y, x)$ — выпуклая по y функция, $\psi(x, x) = 0, \delta > 0$.

Напомним определение обычной μ-сильно выпуклой функции:

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

и определение (δ, L)-модели выпуклой функции f в точке $x \in Q$:

$$\forall y \in Q : f(y) - (f_\delta(x) + \psi(y, x)) \leqslant LV(y, x) + \delta. \tag{4.18}$$

Замечание 4.2.2. Нетрудно заметить, что (δ, L, μ)-модель функции также удовлетворяет (4.18).

Замечание 4.2.3. Если $d(x - y) \leqslant C_n\|x - y\|^2$, $C_n = O(\log n)$, тогда $V(y, x) \leqslant C_n\|x - y\|^2$ и μC_n-сильная выпуклость $\frac{\mu}{2}||y - x||^2 \leqslant f(y) - f(x) - \psi(y, x)$ влечет относительную μ-сильную выпуклость: $\mu V(y, x) + f_\delta(x) + \psi(y, x) \leqslant f(y)$.

Следствие 4.2.4. *Положим в* (4.17) $y = x$, *и тогда* $f_\delta(x) \leqslant f(x) \leqslant f_\delta(x) + \delta$.

Выпишем сначала предложенный А.Д. Агафоновым градиентный метод с неадаптивным выбором шага для задач минимизации функционалов, допускающих существование неточной модели указанного типа.

Алгоритм 11 Неадаптивный градиентный метод для случая (δ, L, μ)-модели целевой функции.

Input: x_0 — начальная точка, $L > 0$, а также $\delta, \widetilde{\delta} > 0$.
for $k \geqslant 0$ **do**
  Set $S_{k+1} := S_k + \frac{1}{L}$.
  $\varphi_{k+1}(x) := \psi(x, x^k) + LV(x, x^k), \quad x^{k+1} := \arg\min_{x \in Q}^{\widetilde{\delta}} \varphi_{k+1}(x).$
end for
Ensure: x^{k+1}

Алгоритм 12 Адаптивный градиентный метод для случае (δ, L, μ)-модели целевой функции.

Input: x^0 — начальная точка, $L_0 > 0$, а также $\delta, \widetilde{\delta} > 0$.
Set $S_0 := 0$
for $k \geqslant 0$ **do**
  Найти наименьшее целое $i_k \geqslant 0$ такое, что

$$f(x^{k+1}) \leqslant f(x^k) + \psi(x^{k+1}, x_k) + L_{k+1}V(x^{k+1}, x^k) + \delta,$$

  где $L_{k+1} = 2^{i_k - 1}L_k$ при $L_k \geqslant 2\mu$ и $L_{k+1} = 2^{i_k}L_k$ при $L_k < 2\mu$.

$$\varphi_{k+1}(x) := \psi(x, x^k) + L_{k+1}V(x, x^k), \quad x^{k+1} := \arg\min_{x \in Q}^{\widetilde{\delta}} \varphi_{k+1}(x).$$

end for
Ensure: x^{k+1}

Теперь приведём предложенный нами метод с адаптивным подбором параметра L на итерациях и доказанную оценку скорости для этого метода.

Напомним следующую вспомогательную лемму (см., например, [22, 168]), основанную на первом равенстве из доказательства теоремы 3.4.4.

Лемма 4.2.5. *Пусть $\psi(x)$ — выпуклая функция и для некоторого $\beta \geqslant 0$*

$$y = \arg\min_{x \in Q}^{\widetilde{\delta}} \{\psi(x) + \beta V(x, z)\}.$$

Тогда для всякого $x \in Q$ (в частности, $x = x_*$) верно неравенство

$$\psi(x) + \beta V(x, z) \geqslant \psi(y) + \beta V(y, z) + \beta V(x, y) - \widetilde{\delta}.$$

Для дальнейших рассуждений введём усредняющий параметр $\widehat{L}$:

$$1 - \frac{\mu}{\widehat{L}} = \sqrt[k+1]{\left(1 - \frac{\mu}{L_{k+1}}\right)\left(1 - \frac{\mu}{L_k}\right)\ldots\left(1 - \frac{\mu}{L_1}\right)}.$$

Обратим внимание, что при $L_i \geq \mu$ $(i = 1, 2, \ldots)$

$$\min_{1 \leqslant i \leqslant k+1} L_i \leqslant \widehat{L} \leqslant \max_{1 \leqslant i \leqslant k+1} L_i \leqslant 2L.$$

Справедливо следующее утверждение.

Теорема 4.2.6. *Пусть $\psi(y, x)$ — (δ, L, μ)-модель для f относительно $V(y, x)$ на Q. Тогда при указанных обозначениях после выполнения k итераций алгоритма 12 верны оценки:*

$$V(x_*, x^{k+1}) \leqslant \frac{2L(\delta + \widetilde{\delta})}{\mu^2}\left(1 - \left(1 - \frac{\mu}{2L}\right)^{k+1}\right) + \left(1 - \frac{\mu}{\widehat{L}}\right)^{k+1} V(x_*, x^0),$$

$$f(x^{k+1}) - f(x_*) \leqslant$$

$$\leqslant \frac{4L^2(\delta + \widetilde{\delta})}{\mu^2}\left(1 - \left(1 - \frac{\mu}{2L}\right)^{k+1}\right) + 2L\left(1 - \frac{\mu}{\widehat{L}}\right)^{k+1} V(x_*, x^0).$$

Доказательство. Пусть алгоритм 12 работает ровно k итераций. Тогда согласно лемме 4.2.5 для всякого $x \in Q$:

$$-\widetilde{\delta} \leqslant \psi_\delta(x, x^k) - \psi_\delta(x^{k+1}, x^k) + L_{k+1}V(x, x^k) - \\ -L_{k+1}V(x, x^{k+1}) - L_{k+1}V(x^{k+1}, x^k).$$

Это означает, что

$$L_{k+1}(x, x^{k+1}) \leqslant \widetilde{\delta} + \psi_\delta(x, x^k) - \psi_\delta(x^{k+1}, x^k) + \\ +L_{k+1}V(x, x^k) - L_{k+1}V(x^{k+1}, x^k). \tag{4.19}$$

Далее, поскольку $\psi(y, x)$ есть (δ, L)-модель относительно $V(y, x)$, то из неравенства

$$f(x^{k+1}) \leqslant f(x^k) + \psi_\delta(x^{k+1}, x^k) + L_{k+1}V(x^{k+1}, x^k) + \delta,$$

получаем

$$-L_{k+1}V(x^{k+1}, x^k) \leqslant \delta - f(x^{k+1}) + f(x^k) + \psi_\delta(x^{k+1}, x^k).$$

Поэтому (4.19) означает, что

$$L_{k+1}V(x, x^{k+1}) \leqslant \widetilde{\delta} + \delta - f(x^{k+1}) + f(x^k) + \psi_\delta(x, x^k) + L_{k+1}V[x^k](x, x^k). \tag{4.20}$$

Поскольку $\psi(y, x)$ есть (δ, L, μ)-модель для f относительно $V(y, x)$, то:

$$f(x^k) + \psi_\delta(x, x^k) \leqslant f(x) - \mu V(x, x^k).$$

Учитываем (4.20), мы получаем:

$$L_{k+1}V(x, x^{k+1}) \leqslant \widetilde{\delta} + \delta + f(x) - f(x^{k+1}) + (L_{k+1} - \mu)V(x, x^k). \tag{4.21}$$

Положим $x = x_*$. Поскольку $L_0 \leqslant 2L$, мы имеем $L_{k+1} \leqslant 2L$ для каждого $k \geq 0$. Также в алгоритме 12, что $L_{k+1} \geq \mu$, и поэтому мы имеем

$$\frac{1}{2L} \leqslant \frac{1}{L_{k+1}} \leqslant \frac{1}{\mu} \quad (\forall k = 0, 1, 2 \ldots).$$

Тогда у нас есть $\forall i \in \mathbb{N} : i < k$

$$\left(1 - \frac{\mu}{L_{k+1}}\right)\left(1 - \frac{\mu}{L_k}\right) \ldots \left(1 - \frac{\mu}{L_{k-i}}\right) \leqslant \left(1 - \frac{\mu}{2L}\right)^{i+1}. \tag{4.22}$$

Таким образом, мы получаем:

$$V(x_*, x^{k+1}) \leqslant \frac{f(x_*) - f(x^{k+1}) + \delta + \widetilde{\delta}}{L_{k+1}} + \left(1 - \frac{\mu}{L_{k+1}}\right) V(x_*, x^k),$$

и

$$\begin{aligned}
&\frac{f(x^{k+1}) - f(x_*)}{L_{k+1}} + V(x_*, x^{k+1}) \leqslant \\
&\leqslant \frac{\delta + \widetilde{\delta}}{L_{k+1}} + \left(1 - \frac{\mu}{L_{k+1}}\right) V(x_*, x^k) \leqslant (\delta + \widetilde{\delta}) \left(\frac{1}{L_{k+1}} + \frac{1}{L_k}\left(1 - \frac{\mu}{L_{k+1}}\right)\right) + \\
&+ \left(1 - \frac{\mu}{L_{k+1}}\right)\left(1 - \frac{\mu}{L_k}\right) V(x_*, x^k) \leqslant \ldots \leqslant \\
&\leqslant (\delta + \widetilde{\delta}) \left(\frac{1}{L_{k+1}} + \frac{1}{L_k}\left(1 - \frac{\mu}{L_k}\right) + \right. \\
&+ \frac{1}{L_{k-1}}\left(1 - \frac{\mu}{L_k}\right)\left(1 - \frac{\mu}{L_{k-1}}\right) + \ldots + \\
&\left. + \frac{1}{L_1}\left(1 - \frac{\mu}{L_k}\right)\left(1 - \frac{\mu}{L_{k-1}}\right)\ldots\left(1 - \frac{\mu}{L_1}\right)\right) + \\
&+ \left(1 - \frac{\mu}{L_{k+1}}\right)\left(1 - \frac{\mu}{L_k}\right)\ldots\left(1 - \frac{\mu}{L_1}\right) V(x_*, x^0).
\end{aligned}$$

Для дальнейшего рассуждения введем средний параметр $\widehat{L}$:

$$1 - \frac{\mu}{\widehat{L}} = \sqrt[k+1]{\left(1 - \frac{\mu}{L_{k+1}}\right)\left(1 - \frac{\mu}{L_k}\right)\ldots\left(1 - \frac{\mu}{L_1}\right)}.$$

Обратите внимание, что ввиду $L_i \geq \mu$ $(i = 1, 2, \ldots)$

$$\min_{1 \leqslant i \leqslant k+1} L_i \leqslant \widehat{L} \leqslant \max_{1 \leqslant i \leqslant k+1} L_i \leqslant 2L.$$

Теперь с учетом (4.22) имеем:

$$\begin{aligned}
&\frac{f(x^{k+1}) - f(x_*)}{L_{k+1}} + V(x_*, x^{k+1}) \leqslant \\
&\leqslant \frac{\delta + \widetilde{\delta}}{\mu} \sum_{i=0}^{k} \left(1 - \frac{\mu}{2L}\right) + \left(1 - \frac{\mu}{\widehat{L}}\right)^{k+1} V(x_*, x^0) \leqslant
\end{aligned} \tag{4.23}$$

$$\leqslant \frac{2L(\delta + \widetilde{\delta})}{\mu^2}\left(1 - \left(1 - \frac{\mu}{2L}\right)^{k+1}\right) + \left(1 - \frac{\mu}{\widehat{L}}\right)^{k+1} V(x_*, x^0). \tag{4.24}$$

Наконец, получаем

$$V(x_*, x^{k+1}) \leqslant \frac{2L(\delta + \widetilde{\delta})}{\mu^2}\left(1 - \left(1 - \frac{\mu}{2L}\right)^{k+1}\right) + \left(1 - \frac{\mu}{\widehat{L}}\right)^{k+1} V(x_*, x^0).$$

и по (4.23)–(4.24) и $L_{k+1} \leqslant 2L$ означает:

$$f(x^{k+1}) - f(x_*) \leqslant$$

$$\leqslant \frac{4L^2(\delta + \widetilde{\delta})}{\mu^2}\left(1 - \left(1 - \frac{\mu}{2L}\right)^{k+1}\right) + 2L\left(1 - \frac{\mu}{\widehat{L}}\right)^{k+1} V(x_*, x^0).$$

□

Обратим внимание, что алгоритм 11 также имеет линейную скорость сходимости для сильно выпуклого случая. Преимущество алгоритма 11 заключается в том, что для его запуска нет необходимости знать параметр сильной выпуклости μ. С другой стороны, этот параметр необходим для оценки качества решения, возвращаемого алгоритмом. Преимущество адаптивного алгоритма 12 состоит в том, что не требуется знания значения параметра L (на него выполняется адаптивная настройка).

4.2.2 Применимость (δ, L, μ)-модели и соответствующих методов к одной задаче описания электоральных процессов

Текст данного подпункта написан совместно со студенткой МФТИ О. А. Кузнецовой (см. [168]; это один из примеров её бакалаврской работы). Рассмотрим пример задачи оптимизации с целевой функцией, которая допускает (δ, L, μ)-модель. Эта задача связана с моделью описания электоральных процессов, предложенной Ю. Е. Нестеровым в [144]. В этой модели избиратели (точки данных) выбирают партию (кластер) итеративным способом путем минимизации следующей функции

$$f_{\mu_1,\mu_2}(x = (z, p)) = g(x) + \mu_1 \sum_{k=1}^{n} z_k \ln z_k + \frac{\mu_2}{2}\|p\|_2^2 \to \min_{z \in S_n(1), p \in \mathbb{R}^m_+}, \quad (4.25)$$

где $\mathbb{R}^m_+$ — неотрицательный ортант и $S_n(1)$ является стандартным n-мерным симплексом в $\mathbb{R}^n$.

Вектор z содержит вероятности, с которыми избиратели выбирают рассматриваемую партию, а вектор p описывает положение партии в

пространстве мнений избирателей. Минимизированный потенциал является результатом объединения двух задач оптимизации в одну: избиратели выбирают партию, позиция которой наиболее близка к их личному мнению, а партия корректирует свою позицию, минимизируя дисперсию и стараясь не отходить слишком далеко от своей первоначальной позиции. Ю.Е. Нестеров в [144] использовал последовательный процесс выборов, чтобы показать, что при некоторых естественных допущениях процесс сближается и дает кластеризацию точек данных. Это было сделано для конкретного выбора функции g, которая имеет ограниченную интерпретируемость. Нужно показать, что предложенная в работе концепция неточной модели целевой функции позволяет построить метод градиентного типа для случая общей функции g, которая не обязательно выпукла. Пусть $g(x)$ — функция (вообще говоря, не выпуклая) с L_g-липшицевым градиентом:

$$\|\nabla g(x) - \nabla g(y)\|_* \leqslant L_g \|x - y\| \quad \forall x, y \in S_n(1) \times \mathbb{R}^m_+,$$

и, следуя [144], числа μ_1, μ_2 выберем так, чтобы $L_g \leqslant \mu_1$ и $L_g \leqslant \mu_2$. Норму $\|\cdot\|$ в $S_n(1) \times \mathbb{R}^m_+$ введём следующим образом $\|(z,p)\|^2 = \|z\|_1^2 + \|p\|_2^2$, где $\|z\|_1 = \sum\limits_{k=1}^{n} z_k$ и $\|p\|_2 = \sqrt{\sum\limits_{k=1}^{m} p_k^2}$.

Пусть $KL(z_x|z_y)$ — расхождение Кульбака–Лейбнера между z_x и z_y. Непосредственно можно проверить (см., например [168]), что для произвольных x, y и z

$$\psi_\delta(x,y) = \langle \nabla g(y), x - y\rangle - L_g \cdot KL(z_x|z_y) - \frac{L_g}{2}\|p_x - p_y\|_2^2 +$$

$$+\mu_1(KL(z_x|\mathbf{1}) - KL(z_y|\mathbf{1})) + \frac{\mu_2}{2}\left(\|p_x\|_2^2 - \|p_y\|_2^2\right)$$

есть $(0, 2L_g)$-модель для $f_{\mu_1,\mu_2}(x)$ в x относительно дивергенции Брегмана вида

$$V(x,y) = KL(z_x|z_y) + \frac{1}{2}\|p_x - p_y\|_2^2.$$

Поэтому $\psi_\delta^{lin}(x,y)$ есть $(0, \max\{\mu_1,\mu_2\} + L_g, \min\{\mu_1,\mu_2\} - L_g)$-модель для функции f_{μ_1,μ_2}, т.е. для произвольных допустимых x и y:

$$f_{\mu_1,\mu_2}(y) + \psi_\delta^{lin}(x,y) + (\min\{\mu_1,\mu_2\} - L_g)V(x,y) \leqslant f_{\mu_1,\mu_2}(x)$$

и

$$f_{\mu_1,\mu_2}(x) \leqslant f_{\mu_1,\mu_2}(y) + \psi_\delta^{lin}(x,y) + (\max\{\mu_1,\mu_2\} + L_g)V(x,y).$$

Это означает, что возможно применить разработанный алгоритм 12 (как и алгоритм 13 из следующего раздела) к задаче (4.25). При этом вспомогательная подзадача на каждой итерации метода будет гладкой и сильно выпуклой относительно выбранной для данного примера дивергенции Брегмана, что гарантирует возможность достижения $\widetilde{\delta}$-решения такой задачи за линейное время.

4.3 Градиентный метод для задач минимизации функционалов, допускающих (δ, Δ, L, μ)-модель функции в произвольной запрошенной точке с адаптивной настройкой параметров

В данном пункте мы рассмотрим аналог введённой выше (δ, L, μ)-модели целевой функции не только с постоянной, но и с переменной неточностью по аналогии с понятием (δ, Δ, L)-модели, которое введено в предыдущей главе работы. В отличие от раздела 4.2 мы обоснуем для такого более общего понятия неточной модели целевого функционала оценку скорости сходимости градиентного метода с адаптивной настройкой не только на константу гладкости L, но на соответствующие возможным погрешностям параметры δ и Δ. При $\mu > 0$ это понятие позволяет обосновать скорость сходимости предлагаемого нами метода, близкую к линейной.

Определение 4.3.1. Будем говорить, что f допускает (δ, Δ, L, μ)-модель $\psi(y,x)$ в точке $x \in Q$, если функционал ψ выпуклый по первой переменной и для произвольного $y \in Q$ верно

$$f(y) \leqslant f_\delta(x) + \psi(y,x) + \Delta\|y - x\| + \delta + LV(y,x), \tag{4.26}$$

а также

$$f(x) - \delta + \psi(x_*,x) + \mu V(x_*,x) \leqslant f_\delta(x) + \psi(x_*,x) + \mu V(x_*,x) \leqslant f(x_*), \tag{4.27}$$

где x_* — ближайшее к x решение задачи минимизации f с точки зрения дивергенции Брэгмана $V(x_*, x)$. Неравенство (4.27) будет, в частности, верно для задачи минимизации сильно квазивыпуклой целевой функции [132] при достаточно малой величине погрешности градиента.

Пример 4.3.2. В качестве примера отметим задачу сильно выпуклой композитной оптимизации $f(x) = g(x) + h(x) \to \min$, где g — гладкая выпуклая функция, а h — выпуклая не обязательно гладкая функция простой структуры. Если при этом для градиента ∇g задано его приближение $\widetilde{\nabla} g$: $\left\|\widetilde{\nabla} g(x) - \nabla g(x)\right\| \leqslant \Delta$, то можно положить

$$\psi(y, x) = \langle \widetilde{\nabla} g(x), y - x\rangle + h(y) - h(x)$$

и в случае μ-сильной выпуклости g или h при подходящем подборе параметров будут выполняться условия определения 4.3.1.

Для задачи минимизации функционала, допускающего (δ, Δ, L, μ)-модель в произвольной запрошенной точке предложим следующий метод.

Алгоритм 13 Адаптивный градиентный метод для функций, допускающих (δ, Δ, L, μ)-модель в запрошенной точке.

Require: $x^0 \in Q$ — начальная точка, $V(x_*, x^0) \leqslant R^2$, параметры $L_0, \Delta_0, \delta_0 > 0:\ 2\mu < L_0 \leqslant 2L,\ \Delta_0 \leqslant 2\Delta,\ \delta_0 \leqslant 2\delta$.

1: $L_{k+1} := \max\{\mu, {}^{L_k}/_2\},\ \Delta_{k+1} := {}^{\Delta_k}/_2,\ \delta_{k+1} := {}^{\delta_k}/_2$.
2: $x^{k+1} := \arg\min\limits_{x \in Q}\{\psi(x, x^k) + L_{k+1} V(x, x^k)\}$.
3: **repeat**
4: **if** $f_\delta(x^{k+1}) \leqslant f_\delta(x^k) + \psi(x^{k+1}, x^k) + L_{k+1} V(x^{k+1}, x^k) + \delta_{k+1} + \Delta_{k+1}\left\|x^{k+1} - x^k\right\|$ **then**
5: $k := k + 1$ и выполнение п. 1.
6: **else**
7: $L_{k+1} := 2L_{k+1}\ \delta_{k+1} := 2 \cdot \delta_{k+1};\ \Delta_{k+1} := 2 \cdot \Delta_{k+1}$ и выполнение п. 2.
8: **end if**
9: **until** $k \geqslant N$

Ensure: $y^{k+1} := \arg\min\limits_{i=0,\ldots,k} f(x^{i+1})$.

Далее, условимся полагать, что $\prod\limits_{i=k+1}^{k} a_i = 1$ для некоторой числовой последовательности a_i. Справедлива следующая

Теорема 4.3.3. *Пусть f имеет (δ, Δ, L, μ)-модель в каждой точке $x \in Q$. Тогда после k итераций справедливо неравенство*

$$f(y^{k+1}) - f(x_*) \leqslant \frac{1}{\sum\limits_{i=0}^{k} \frac{1}{L_{i+1}} \prod\limits_{j=i+1}^{k} \left(1 - \frac{\mu}{L_j}\right)} \left(\prod_{i=0}^{k} \left(1 - \frac{\mu}{L_{i+1}}\right) V(x_*, x^0) + \right.$$

$$\left. + \sum_{i=0}^{k} \frac{\delta + \delta_{i+1} + \Delta_{i+1} \left\| x^{i+1} - x^i \right\|}{L_{i+1}} \prod_{j=i+1}^{k} \left(1 - \frac{\mu}{L_j}\right) \right).$$

Отметим, что вспомогательная задача п. 2 листинга алгоритма 13 решается не более

$$2k + \max \left\{ \log_2 \frac{2L}{L_0}, \log_2 \frac{2\delta}{\delta_0}, \log_2 \frac{2\Delta}{\Delta_0} \right\} \tag{4.28}$$

раз.

Доказательство. Введем обозначения:

$$\widehat{\delta}_{k+1} := \delta_{k+1} + \Delta_{k+1} \left\| x^{k+1} - x^k \right\|.$$

После k итераций алгоритма 13 согласно лемме 4.2.5 получаем

$$0 \leqslant \psi(x, x^k) - \psi(x^{k+1}, x^k) + L_{k+1} V(x, x^k) -$$

$$- L_{k+1} V(x, x^{k+1}) - L_{k+1} V(x^{k+1}, x^k),$$

откуда

$$L_{k+1} V(x, x^{k+1}) \leqslant \psi(x, x^k) - \psi(x^{k+1}, x^k) + L_{k+1} V(x, x^k) - L_{k+1} V(x^{k+1}, x^k). \tag{4.29}$$

Согласно неравенству (4.26):

$$-L_{k+1} V(x^{k+1}, x^k) \leqslant \widehat{\delta}_{k+1} - f_\delta(x^{k+1}) + f_\delta(x^k) + \psi(x^{k+1}, x^k).$$

Применяя теперь (4.29), получаем

$$L_{k+1} V(x, x^{k+1}) \leqslant \widehat{\delta}_{k+1} - f(x^{k+1}) + f_\delta(x^k) + \psi(x, x^k) + L_{k+1} V(x, x^k) + \delta. \tag{4.30}$$

Пусть $x = x_*$. Тогда, учитывая (4.27), имеем $f_\delta(x^k)+\psi(x_*, x^k) \leqslant f(x_*)-\mu V(x_*, x^k)$. Применим это неравенство к (4.30): $L_{k+1}V(x_*, x^{k+1}) \leqslant \delta + \widehat{\delta}_{k+1} + f(x_*) - f(x^{k+1}) + (L_{k+1} - \mu)V(x_*, x^k)$. Далее,

$$V(x_*, x^{k+1}) \leqslant \frac{f(x_*) - f(x^{k+1})}{L_{k+1}} + \frac{\delta + \widehat{\delta}_{k+1}}{L_{k+1}} + \left(1 - \frac{\mu}{L_{k+1}}\right) V(x_*, x^k) \leqslant$$

$$\leqslant \frac{f(x_*) - f(x^{k+1})}{L_{k+1}} + \frac{1}{L_k}\left(1 - \frac{\mu}{L_{k+1}}\right)\left(f(x_*) - f(x^k)\right) + \frac{\delta + \widehat{\delta}_{k+1}}{L_{k+1}} +$$

$$+ \frac{\delta + \widehat{\delta}_k}{L_k}\left(1 - \frac{\mu}{L_{k+1}}\right) + \left(1 - \frac{\mu}{L_{k+1}}\right)\left(1 - \frac{\mu}{L_k}\right) V(x_*, x^{k-1}) \leqslant$$

$$\leqslant \sum_{i=0}^{k} \frac{f(x_*) - f(x^{i+1})}{L_{i+1}} \prod_{j=i+1}^{k}\left(1 - \frac{\mu}{L_j}\right) + \sum_{i=0}^{k} \frac{\delta + \widehat{\delta}_{i+1}}{L_{i+1}} \prod_{j=i+1}^{k}\left(1 - \frac{\mu}{L_j}\right) +$$

$$+ \prod_{i=0}^{k}\left(1 - \frac{\mu}{L_{i+1}}\right) V(x_*, x_0).$$

С учетом $y^{k+1} = \arg\min\limits_{i=\overline{0,k}} f(x^{i+1})$ и $V(x_*, x^{k+1}) \geqslant 0$ имеем

$$f(y^{k+1}) - f(x_*) \leqslant \frac{1}{\sum\limits_{i=0}^{k} \frac{1}{L_{i+1}} \prod\limits_{j=i+1}^{k}\left(1 - \frac{\mu}{L_j}\right)} \left(\prod_{i=0}^{k}\left(1 - \frac{\mu}{L_{i+1}}\right) V(x_*, x^0) + \right.$$

$$\left. + \sum_{i=0}^{k} \frac{\delta + \widehat{\delta}_{i+1}}{L_{i+1}} \prod_{j=i+1}^{k}\left(1 - \frac{\mu}{L_j}\right)\right).$$

Оценка (4.28) обосновывается аналогично п. 2 доказательства теоремы 3.1.4.

□

Замечание 4.3.4. Если убрать соответствующее ограничение пункта 1 листинга алгоритма 13 и допустить возможность при некотором $k \geqslant 0$ $L_{k+1} < \mu$, то в таком случае на этой итерации будет верно $f(x^{k+1}) - f(x_*) \leqslant \delta + \widehat{\delta}_{k+1}$.

Замечание 4.3.5. Оценка (4.28) показывает, что в среднем трудоемкость итерации предложенного адаптивного алгоритма превышает

трудоемкость неадаптивного метода не более, чем в постоянное число раз. Отметим также, что при $k = 0, 1, 2, \ldots$ $L_{k+1} \leqslant 2CL$, $C = \max\left\{1, \frac{2\delta}{\delta_0}, \frac{2\Delta}{\Delta_0}\right\}$. Это указывает на линейную скорость сходимости рассматриваемого метода в случае отсутствия погрешностей. Для малых величин параметров, соответствующих погрешностям, предыдущая теорема позволяет гарантировать скорость сходимости, близкую к линейной.

Следствие 4.3.6. *При $\mu = 0$ полученная оценка качества решения принимает вид:*

$$f(y^{k+1}) - f(x_*) \leqslant$$

$$\leqslant \frac{V(x_*, x^0)}{\sum\limits_{i=0}^{k} \frac{1}{L_{i+1}}} + \left(\sum_{i=0}^{k} \frac{1}{L_{i+1}}\right)^{-1} \sum_{i=0}^{k} \frac{\delta_{i+1} + \Delta_{i+1} \left\|x^{i+1} - x^i\right\|}{L_{i+1}} + \delta \leqslant$$

$$\leqslant 2CLV(x_*, x^0) + \left(\sum_{i=0}^{k} \frac{1}{L_{i+1}}\right)^{-1} \sum_{i=0}^{k} \frac{\delta_{i+1} + \Delta_{i+1} \left\|x^{i+1} - x^i\right\|}{L_{i+1}} + \delta.$$

Покажем, как можно получить для (δ, Δ, L, μ)-модели аналог результатов раздела 3.3. Если положить $\delta_{i+1} = \delta = 0$ для всякого $i = \overline{0, k}$ и $\Delta_{i+1} \left\|x^{i+1} - x^i\right\| \leqslant \frac{\varepsilon}{2}$, то согласно теореме 4.3.3

$$f(y^{k+1}) - f(x_*) \leqslant 2L \left(1 - \frac{\mu}{2L}\right)^{k+1} V(x_*, x^0) + \frac{\varepsilon}{2}.$$

В частности, данное неравенство позволяет при понимании величины Δ в (4.26) как характеристики негладкости функционала f применить процедуру, аналогичную (3.14) (при этом уже необходимо требовать 1-сильную выпуклость прокс-функции в определении 4.3.1). Тогда ε-точность решения задачи минимизации f будет достигаться за

$$O\left(\frac{1}{\varepsilon} \log_2^2 \frac{1}{\varepsilon}\right)$$

шагов градиентного метода (пункта 2 листинга алгоритма 13). Действительно, если положить $\widehat{\delta} = \delta = 0$ (то есть $f_\delta = f$) и $V(x_*, x^0) \leqslant R^2$, то для достижения качества решения

$$f(y^{k+1}) - f(x_*) \leqslant \varepsilon$$

необходимо выполнить не более

$$\left\lceil \frac{2L}{\mu} \ln \frac{4LR^2}{\varepsilon} \right\rceil - 1$$

итераций алгоритма 13. По аналогии с рассуждениями раздела 3.3 (доказательство теоремы 3.3.9) ввиду оценки (3.17) с учётом процедуры (3.14) на каждом шаге итоговое число обращений к (суб)градиенту целевого функционала для алгоритма 13 можно оценить как

$$\left\lceil \left(\frac{2L}{\mu} + \frac{32C\Delta^2}{\mu\varepsilon} \right) \ln \left(\frac{4LR^2}{\varepsilon} + \frac{64C\Delta^2R^2}{\varepsilon^2} \right) \right\rceil \left\lceil \log_2 \left(1 + \frac{16\Delta^2}{\varepsilon L} \right) \right\rceil - 1. \tag{4.31}$$

Заметим, что в данном пункте мы опускаем вопрос влияния на итоговые оценки погрешности $\widetilde{\delta}$ решения вспомогательных задач на итерациях методов (см. также метод для вариационных неравенств из раздела 3.4). Это связано с тем, что в отличие от погрешности оракула (модели) δ, по-видимому, не удаётся реализовать адаптивную настройку на величину погрешностей для вспомогательных задач $\widetilde{\delta}$. Для фиксированных величин таких погрешностей оценки скорости сходимости градиентных методов получены в предыдущем пункте. В данном пункте мы сфокусировались уже на новом идее, связанной со способом адаптивной настройки не только на константы (уровень) гладкости L, а и на величины погрешностей модели δ и Δ.

Замечание 4.3.7. По аналогии с рассуждениями раздела 3.3 обсудим вариант использования ускоренного варианта градиентного метода для задач, допускающих существование (δ, Δ, L, μ)-модели целевой функции в произвольной запрошенной точке. Мы отправляемся от алгоритма 2 из [178], но уже для предлагаемого нами варианта понятия абстрактной оптимизационной модели с двумя параметрами δ и Δ, соответствующими погрешностям (алгоритм 14) при $f_\delta = f$ при условии

$$f_\delta(x) + \psi(y,x) + \frac{\mu}{2}\|y-x\|^2 \leqslant f(y) \leqslant$$

$$\leqslant f_\delta(x) + \psi(y,x) + \Delta\|y-x\| + \delta + \frac{L}{2}\|y-x\|^2 \leqslant$$

$$\leqslant f_\delta(x) + \psi(y,x) + \Delta\|y-x\| + \delta + LV(y,x) \quad \forall x, y \in Q.$$

В отличие от рассуждений раздела 3.3 ограничимся лишь неадаптивным подходом и выберем постоянный параметр $L_{k+1} = 2^p L$ для некоторого натурального p. Покажем, что возможно найти такое натуральное

Алгоритм 14 Быстрый градиентный метод для функций, допускающих (δ, Δ, L, μ)-модель в произвольной запрошенной точке

Input: x^0 — начальная точка, $\mu \geqslant 0$, $L_0 > 0$, δ_0, $\Delta_0 > 0$.

Пусть $y^0 := x^0$, $u^0 := x^0$, $\alpha_0 := 0$, $A_0 := \alpha_0$

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

Найти наименьшее целое $i_k \geqslant 0$, такое, что

$$f_\delta(x^{k+1}) \leqslant f_\delta(y^{k+1}) + \\ + \psi(x^{k+1}, y^{k+1}) + \frac{L_{k+1}}{2} \left\|x^{k+1} - y^{k+1}\right\|^2 + \\ + \Delta_{k+1} \left\|x^{k+1} - y^{k+1}\right\| + \delta_{k+1},$$

где $L_{k+1} = \max\left\{\mu, 2^{i_k - 1} L_k\right\}$, $\Delta_{k+1} = 2^{i_k - 1}\Delta_k$, $\delta_{k+1} = 2^{i_k - 1}\delta_k$, α_{k+1} — наибольший корень уравнения

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

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

$$\varphi_{k+1}(x) = \alpha_{k+1}\psi_{\delta_k}(x, y^{k+1}) + (1 + A_k\mu)V[u^k](x) + \alpha_{k+1}\mu V[y^{k+1}](x).$$

$$u^{k+1} := \arg\min_{x \in Q}^{\widetilde{\delta}_k} \varphi_{k+1}(x).$$

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

end for

p, для которого после завершения k-ой итерации гарантированно будет выполнено неравенство

$$f(x^{k+1}) \leqslant f(y^{k+1}) + \psi(x^{k+1}, y^{k+1}) + \frac{2^p L}{2} \left\|x^{k+1} - y^{k+1}\right\|^2 + \frac{\varepsilon}{2\gamma}.$$

В таком случае известно [91, 178], что после N итераций справедливо неравенство

$$f(x^N) - f^* \leqslant 2^{p+1} L R^2 \exp\left(-\frac{N-1}{2}\sqrt{\frac{\mu}{2^p L}}\right) + \left(1 + \sqrt{\frac{2^p L}{\mu}}\right)\delta. \quad (4.32)$$

Будем подбирать p так, чтобы гарантированно была верна альтернатива $\forall k = \overline{1, N}$

$$\left[\begin{array}{c} \Delta \left\|x^{k+1} - y^{k+1}\right\| \leqslant \dfrac{\varepsilon}{2\gamma}, \\ \dfrac{(2^p - 1)L}{2} \left\|x^{k+1} - y^{k+1}\right\|^2 \geqslant \Delta \left\|x^{k+1} - y^{k+1}\right\|. \end{array}\right.$$

Если $\Delta \left\|x^{k+1} - y^{k+1}\right\| > \frac{\varepsilon}{2\gamma}$, то $\frac{(2^p-1)L}{2} \left\|x^{k+1} - y^{k+1}\right\| > \frac{(2^p-1)L\varepsilon}{4\gamma\Delta}$. Поэтому можно выбрать p так, чтобы

$$2^p > 1 + \frac{4\gamma\Delta^2}{L\varepsilon}.$$

Выберем в (4.32) количество шагов N так, чтобы для некоторого γ было верно

$$\begin{cases} 2^{p+1} L R^2 \exp\left(-\frac{N-1}{2}\sqrt{\frac{\mu}{2^p L}}\right) \leqslant \frac{\varepsilon}{2}, \\ \left(1 + \sqrt{\frac{2^p L}{\mu}}\right)\frac{\varepsilon}{2\gamma} \leqslant \frac{\varepsilon}{2}. \end{cases} \quad (4.33)$$

Ясно, что выполнение условий (4.33) гарантирует достижение качества решения по функции $f(x^N) - f^* \leqslant \varepsilon$. Второе неравенство в (4.33) означает, что

$$\gamma \geqslant 1 + \sqrt{\frac{2^p L}{\mu}} > 1 + \sqrt{\left(1 + \frac{4\gamma\Delta^2}{L\varepsilon}\right)\frac{L}{\mu}},$$

откуда $\gamma > 1$ и

$$\gamma^2 - 2\gamma\left(1 + \frac{2\Delta^2}{\varepsilon\mu}\right) + 1 - \frac{L}{\mu} > 0.$$

Это означает, что

$$\gamma > 1 + \frac{2\Delta^2}{\varepsilon\mu} + \sqrt{\frac{4\Delta^4}{\varepsilon^2\mu^2} + \frac{4\Delta^2}{\varepsilon\mu} + \frac{L}{\mu}} \geqslant$$

$$\geqslant 1 + \left(2 + \frac{2}{\sqrt{3}}\right)\frac{\Delta^2}{\varepsilon\mu} + \frac{2\Delta}{\sqrt{3\varepsilon\mu}} + \sqrt{\frac{L}{3\mu}}.$$

Теперь из первого неравенства в (4.33) оценим необходимое количество итераций N:

$$\exp\left(-\frac{N-1}{2}\sqrt{\frac{\mu}{2^p L}}\right) \leqslant \frac{\varepsilon}{2^{p+1}LR^2},$$

откуда

$$N \geqslant 1 + \sqrt{\frac{2^p L}{\mu}}\ln\frac{2^{p+1}LR^2}{\varepsilon} >$$

$$> 1 + \sqrt{\frac{L}{\mu}\left(1 + \frac{4\gamma\Delta^2}{L\varepsilon}\right)}\ln\left(\left(1 + \frac{4\gamma\Delta^2}{L\varepsilon}\right)2LR^2\right) =$$

$$= 1 + \sqrt{\frac{L}{\mu}\left(1 + \frac{4\gamma\Delta^2}{L\varepsilon}\right)}\ln\left(2LR^2 + \frac{8\gamma\Delta^2R^2}{\varepsilon}\right) >$$

$$> 1 + \sqrt{\frac{L}{\mu} + \frac{4\Delta^2}{\mu\varepsilon}\left(1 + \left(2 + \frac{2}{\sqrt{3}}\right)\frac{\Delta^2}{\varepsilon\mu} + \frac{2\Delta}{\sqrt{3\varepsilon\mu}} + \sqrt{\frac{L}{3\mu}}\right)}\cdot$$

$$\cdot\ln\left(2LR^2 + \frac{8\Delta^2R^2}{\varepsilon}\left(1 + \frac{2\Delta^2}{\varepsilon\mu} + \frac{2\Delta^2}{\varepsilon\mu\sqrt{3}} + \frac{2\Delta}{\sqrt{3\varepsilon\mu}} + \sqrt{\frac{L}{3\mu}}\right)\right) =$$

$$= 1 + \sqrt{\frac{L}{\mu} + \left(\frac{8}{\sqrt{3}} + 8\right)\frac{\Delta^4}{\varepsilon^2\mu^2} + \frac{8\Delta^3}{\mu\varepsilon\sqrt{3\mu\varepsilon}} + \frac{4\Delta^2}{\mu\varepsilon}\sqrt{\frac{L}{3\mu}}}\cdot$$

$$\cdot\ln\left(2LR^2 + \frac{8\Delta^2R^2}{\varepsilon}\left(1 + \left(2 + \frac{2}{\sqrt{3}}\right)\frac{\Delta^2}{\varepsilon\mu} + \frac{2\Delta}{\sqrt{3\varepsilon\mu}} + \sqrt{\frac{L}{3\mu}}\right)\right).$$

Поэтому можно считать необходимое количество итераций сопоставимым с

$$\left(O\left(\sqrt{\frac{L}{\mu}}\right) + O\left(\frac{\Delta^2}{\varepsilon\mu}\right) + O\left(\left(\frac{\Delta^2}{\varepsilon\mu}\right)^{3/2}\right) + O\left(\frac{2\Delta\sqrt[4]{L}}{\sqrt[4]{\mu^3}\varepsilon^2}\right)\right)\cdot \qquad (4.34)$$

$$\cdot \ln \left(2LR^2 + \frac{8\Delta^2 R^2}{\varepsilon} \left(1 + \left(2 + \frac{2}{\sqrt{3}} \right) \frac{\Delta^2}{\varepsilon\mu} + \frac{2\Delta}{\sqrt{3\varepsilon\mu}} + \sqrt{\frac{L}{3\mu}} \right) \right).$$

В отличие от аналогичных результатов в разделе 3.3 (случай $\mu = 0$) уже не очевидно, что (4.34) лучше (4.31) (для неускоренного метода) при $\Delta > 0$. Тем не менее, при $\Delta = 0$ оценка (4.34) имеет вид $O\left(\sqrt{\frac{L}{\mu}} \ln \frac{LR^2}{\varepsilon}\right)$, что может быть лучше оценки (4.31), так как $\sqrt{\frac{L}{\mu}} \leqslant \frac{L}{\mu}$.

4.4 Адаптивный метод для задач сильно выпуклого программирования с одним ограничением

Рассмотрим задачу вида

$$f(x) \to \min, \quad x \in Q \subset \mathbb{R}^n,\ g(x) \leqslant 0, \tag{4.35}$$

где Q — выпуклый компакт в конечномерном нормированном пространстве $\mathbb{R}^n$, f и g — выпуклые функционалы, g удовлетворяет условию Липшица (относительно евклидовой нормы)

$$|g(y) - g(x)| \leqslant M_g \, ||x - y||_2, \ \ \forall x, y \in Q$$

при некоторой постоянной $M_g > 0$. При этом если функционалов ограничений несколько $\{g_p(x)\}_{p=1}^m$, то можно рассмотреть задачу с одним ограничением $g(x) = \max\limits_{p=\overline{1,m}} g_p(x)$, которое будет заведомо удовлетворять условию Липшица, если все g_p удовлетворяют условию Липшица. Вполне естественно рассматривать подход, основанный на замене (4.35) двойственной к ней задачей

$$\varphi(\lambda) = \min_{x \in Q} \{ f(x) + \lambda g(x) \} \to \max_{\lambda \geqslant 0}. \tag{4.36}$$

В этом случае двойственная функция зависит от одной двойственной переменной $\lambda \geqslant 0$. Если выполнены условия Слейтера для задачи (4.35), то возможные значения λ ограничены отрезком. Это позволяет применять метод дихотомии аналогично для поиска значения двойственной переменной λ, которая близка к соответствующей переменной λ_*, для которой

$$\lambda_* \cdot g(x(\lambda_*)) = 0.$$

Однако для эффективного решения (4.36) необходимо решать вспомогательную задачу многомерной минимизации по x функционала $f(x) + \lambda g(x)$ при фиксированном λ. Вообще говоря, такая задача может быть решена лишь с некоторой точностью методами оптимизации. Это приводит к погрешностям при нахождении $\varphi(\lambda)$ и ее производной $\varphi'(\lambda)$. Также если $f(x) + \lambda g(x)$ не сильно выпукла, то φ может быть негладкой в точке λ. Поэтому при рассмотрении указанного подхода вполне естественно потребовать сильную выпуклость целевого функционала f.

В данном пункте мы рассмотрим алгоритм для задач вида (4.35) с сильно выпуклым целевым функционалом f при следующих типах предположений для f и g:

$$|f(x) - f(y)| \leqslant M_f ||x - y||_2, \quad |g(x) - g(y)| \leqslant M_g ||x - y||_2 \tag{4.37}$$

или

$$||\nabla f(x) - \nabla f(y)||_2 \leqslant L_f ||x - y||_2, \quad ||\nabla g(x) - \nabla g(y)||_2 \leqslant L_g ||x - y||_2 \tag{4.38}$$

для всех $x, y \in Q$ и для некоторых действительных положительных чисел M_f, M_g, L_f, L_g. Докажем, что

- При условиях (4.38) предлагаемый метод имеет сложность (необходимое количество итераций для достижения ожидаемого качества решения) не более

$$O\left(\log_2^2 \frac{1}{\varepsilon}\right);$$

- В предположениях (4.37) показано, что сложность предложенного метода не превышает

$$O\left(\frac{1}{\varepsilon^2} \log_2 \frac{1}{\varepsilon}\right),$$

что нельзя считать оптимальным с точки зрения нижних оценок. Однако адаптивность предложенного метода позволяет улучшить скорость его работы по сравнению с теоретическими оценками, что проиллюстрировано некоторыми численными экспериментами для примеров негладких задач.

Напомним, что всюду в работе под сложностью предложенного мы понимаем зависимость от ε количества необходимых итераций (обращений к оракулу, выдающему значение (суб)градиента некоторой функции в запрошенной точке) для достижения ε-точного решения задачи.

Отметим, что возможно рассматривать задачу (4.36) как седловую и применять к ней методы, эффективно работающие для седловых задач. Если понимать эффективность в терминах нижних оракульных оценок, то одним из наиболее известных оптимальных методов как для гладких, так и для негладких выпукло-вогнутых седловых задач является экстраградиентный метод, а также его современный аналог — проксимальный зеркальный метод А. С. Немировского. В главе 2 мы описали предложенный недавно универсальный аналог этого метода (UMP), который предполагает адаптивную настройку на уровень гладкости седловой задачи. Для гладкого случая (в предположении (4.38)) универсальный метод из главы 2 будет сходится со скоростью не ниже $O\left(\varepsilon^{-1}\right)$, а для негладкого (в предположении (4.37)) — со скоростью не ниже $O\left(\varepsilon^{-2}\right)$. Однако за счёт адаптивности метода возможно существенное ускорение по сравнению с теоретическими оценками. Как видим, наш подход (алгоритм 15) за счёт сужения класса задач в гладком случае приводит к лучшей оценке скорости сходимости, а в негладком — к худшей. В завершении данного раздела приведены результаты экспериментов, которые показывают, что за счёт адаптивности метода ситуация может оказаться иной: алгоритм 15 может работать существенно лучше обоснованной нами теоретической оценки и лучше рассмотренного в разделе 2 универсального метода для вариационных неравенств и седловых задач.

4.4.1 Постановка задачи и вспомогательные результаты

Пусть Q — компактное подмножество конечномерного нормированного векторного пространства со скалярным произведением $\langle \cdot, \cdot \rangle$ и со стандартной евклидовой нормой $||x||_2 = \sqrt{\langle x, x \rangle}$. Рассмотрим следующую оптимизационную задачу

$$\min_{\substack{g(x) \leqslant 0 \\ x \in Q}} f(x), \tag{4.39}$$

где f — выпуклая функция и g — μ_g-сильно выпуклая функция в 2-норме, т.е.

$$g(\alpha x + (1-\alpha)y) \leqslant \alpha g(x) + (1-\alpha)g(y) - \alpha(1-\alpha)\frac{\mu_g}{2}||x-y||_2^2$$

для $\alpha \in [0,1]$ и для всех $x, y \in Q$. Предположим, что f и g удовлетворяют условию Липшица, т.е.

$$|f(y) - f(x)| \leqslant M_f \, ||y-x||_2 \quad \forall x, y \in Q,$$

$$|g(y) - g(x)| \leqslant M_g \, ||y-x||_2 \quad \forall x, y \in Q.$$

Введем двойственный множитель $\lambda \geqslant 0$ и запишем двойственную задачу к задаче (4.39)

$$\min_{\substack{g(x)\leqslant 0 \\ x\in Q}} f(x) = \min_{x\in Q}\left\{f(x) + \max_{\lambda\geqslant 0}(\lambda g(x))\right\} = \max_{\lambda\geqslant 0}\Big\{\underbrace{\min_{x\in Q}\left(f(x)+\lambda g(x)\right)}_{=\varphi(\lambda)}\Big\}.$$

То есть двойственная задача к задаче (4.39) имеет следующий вид

$$\varphi(\lambda) = f(x(\lambda)) + \lambda g(x(\lambda)) \to \max_{\lambda\geqslant 0}, \tag{4.40}$$

где

$$x(\lambda) = \arg\min_{x\in Q}\left\{f(x) + \lambda g(x)\right\}. \tag{4.41}$$

В дальнейшем нам понадобится известная теорема Демьянова-Данскина-Рубинова, см. [23, 24].

Лемма 4.4.1. *Пусть для любого $\lambda \geqslant 0$ выполняется: $\varphi(\lambda) = \min\limits_{x\in X} F(x, \lambda)$, $F(x, \lambda)$ — выпуклая и гладкая по λ функция и максимум достигается в единственной точке $x(\lambda)$. Тогда $\varphi'(\lambda) = F'_\lambda(x(\lambda), \lambda)$.*

В нашем случае для задачи (4.40) из утверждения 2 получаем

$$\varphi'(\lambda) = g(x(\lambda)). \tag{4.42}$$

Пусть λ^* — решение двойственной задачи (4.40). Тогда, согласно необходимому условию экстремума, для λ^* должно выполняться условие дополняющей нежёсткости

$$\lambda^* g(x(\lambda^*)) = 0, \ \lambda^* \geqslant 0,$$

что, учитывая соотношение (4.42), можно представить в следующем виде

$$\lambda^* \varphi'(\lambda^*) = 0, \ \lambda^* \geqslant 0.$$

4.4.2 Алгоритм В данном параграфе рассмотрим алгоритм для решения задачи, описанной выше. Для решения одномерной двойственной задачи предлагается использовать метод дихотомии (деления отрезка пополам). Поскольку рассматриваются задачи с одним ограничением, то двойственная задача одномерна и возможно применить к ней методы одномерной оптимизации, например, метод дихотомии. Пусть $x_\delta(\lambda)$:

$$f(x_\delta(\lambda)) + \lambda g(x_\delta(\lambda)) - f(x(\lambda)) - \lambda g(x(\lambda)) \leqslant \delta. \tag{4.43}$$

Алгоритм 15

Require: μ-сильно выпуклая функция f, промежуток локализации $\left[\lambda^0_{min}, \lambda^0_{max}\right]$ двойственной переменной, точность решения вспомогательных задач $\delta > 0$, точность ε для (4.36).

1: $N := 0$
2: **repeat**
3: $\quad \lambda^N := \frac{\lambda^N_{min} + \lambda^N_{max}}{2}$;
4: $\quad x_\delta(\lambda^N) = \arg\min\limits_{x \in Q}\{f(x) + \lambda^N g(x)\}$;
5: $\quad \varphi'(\lambda^N) = g(x_\delta(\lambda^N))$;
6: $\quad$ **if** $\varphi'(\lambda^N) < 0$ **then** $\lambda^{N+1}_{max} := \frac{\lambda^N_{min} + \lambda^N_{max}}{2}$;
7: $\quad$ **if** $\varphi'(\lambda^N) > 0$ **then** $\lambda^{N+1}_{min} := \frac{\lambda^N_{min} + \lambda^N_{max}}{2}$;
8: $\quad N := N + 1$;
9: **until** $\lambda^N |g(x_\delta(\lambda^N))| \leqslant \varepsilon$.

Ensure: λ^N,: $\lambda^N |g(x_\delta(\lambda^N))| \leqslant \varepsilon$; $x_\delta(\lambda^N)$.

Для того, чтобы использовать метод дихотомии при решении двойственной задачи, необходимо локализовать значения двойственной переменной. Поскольку мы имеем ограничение-неравенство, то в качестве нижней границы промежутка значений двойственной переменной можно выбрать $\lambda_{min} = 0$. Для определения верхней границы этого промежутка будем использовать условие Слейтера. Напомним следующее утверждение.

Лемма 4.4.2. *Рассмотрим задачу выпуклой оптимизации*

$$f(x) \to \min_{\substack{g(x) \leqslant 0 \\ x \in Q}}.$$

Предположим, что условие Слейтера выполнено, тогда существует такая точка $\overline{x} \in Q$, *для которой выполняется* $g(\overline{x}) < 0$, *т.е.* $g(\overline{x}) = -\gamma < 0$. *В этом случае справедлива следующая оценка:*

$$\lambda^* \leqslant \frac{1}{\gamma}\Big(f(\overline{x}) - \min_{x \in Q} f(x)\Big),$$

где λ^* *— решение двойственной задачи* $\varphi(\lambda) \to \max\limits_{\lambda \geqslant 0}$.

Таким образом, получаем, что в качестве верхней границы для двойственной переменной можно выбрать $\lambda_{max} = \dfrac{1}{\gamma}\Big(f(\overline{x}) - \min\limits_{x \in Q} f(x)\Big)$.

4.4.3 Оценка скорости сходимости метода дихотомии

Теорема 4.4.3. *Пусть* $f(x)$ *—* μ_f*-сильно выпуклая функция, а функция* $g(x)$ *— выпуклая функция и удовлетворяет условию Липшица с константой* M_g. *Тогда для производной функции (4.40), где* $x(\lambda)$ *определяется условием* (4.41), *имеет место следующая оценка:*

$$|\varphi'(\lambda_2) - \varphi'(\lambda_1)| \leqslant \frac{M_g^2}{\mu_f} |\lambda_2 - \lambda_1| .$$

Доказательство. Пусть $\lambda_1,\ \lambda_2 \in [\lambda_{min}, \lambda_{max}]$. Определим

$$x_1 = \arg\min_{x \in Q} \{f(x) + \lambda_1 g(x)\}, \qquad x_2 = \arg\min_{x \in Q} \{f(x) + \lambda_2 g(x)\}.$$

Так как $f(x)$ — μ_f-сильно выпуклая функция, а $g(x)$ — выпуклая функция, то $F_\lambda(x) = f(x) + \lambda g(x)$ является μ_f-сильно выпуклой функцией. Поскольку ввиду сильной выпуклости целевой функции x_1 и x_2 — единственные решения вспомогательных задач, то используя лемму 4.4.1, получаем $\varphi'(\lambda_1) = g(x_1)$, $\varphi'(\lambda_2) = g(x_2)$. В силу сильной выпуклости $F_\lambda(x)$ справедливы следующие оценки:

$$\frac{\mu_f}{2}||x_1 - x_2||_2^2 \leqslant f(x_2) + \lambda_1 g(x_2) - f(x_1) - \lambda_1 g(x_1),$$

$$\frac{\mu_f}{2}||x_1 - x_2||_2^2 \leqslant f(x_1) + \lambda_2 g(x_1) - f(x_2) - \lambda_2 g(x_2).$$

Суммируя данные неравенства, получаем

$$\mu_f||x_1 - x_2||_2^2 \leqslant (\lambda_1 - \lambda_2)\,(g(x_2) - g(x_1)) \leqslant |\lambda_1 - \lambda_2|\,|g(x_2) - g(x_1)| \leqslant$$

$$\leqslant M_g\,|\lambda_1-\lambda_2|\,||x_1-x_2||,$$

так как

$$|g(x)-g(y)|\leqslant M_g||x-y||,\ \ \forall\, x,\, y\in Q.$$

Тогда при $x_1\neq x_2$

$$\|x_2-x_1\|_2\leqslant\frac{M_g}{\mu_f}|\lambda_1-\lambda_2|.$$

В результате получаем

$$|\varphi'(\lambda_2)-\varphi'(\lambda_1)|=|g(x_2)-g(x_1)|\leqslant M_g\|x_2-x_1\|_2\leqslant\frac{M_g^2}{\mu_f}|\lambda_1-\lambda_2|.$$

4.4.4 Алгоритм и оценки скорости его сходимости

Рассмотрим следующий метод для решения поставленной задачи (алгоритм 15). Для вывода оценки скорости сходимости метода понадобится следующее вспомогательное утверждение.

Лемма 4.4.4. *Если выполнен критерий остановки алгоритма 15 при $\lambda=\ \lambda^N$, то справедливы следующие неравенства:*

$$f(x_\delta(\lambda))-f(x_*)\leqslant\varepsilon+\delta,\quad g(x_\delta(\lambda))\leqslant\frac{\varepsilon}{\lambda},$$

где δ — точность решения по функции вспомогательной задачи в п. 4 листинга алгоритма 15.

В частности, в случае $\delta=\varepsilon$ справедливы такие неравенства:

$$f(x_\delta(\lambda))-f(x_*)\leqslant 2\varepsilon,\quad g(x_\delta(\lambda))\leqslant\frac{\varepsilon}{\lambda}.$$

Доказательство. Пусть λ^* — решение двойственной задачи (4.40). Обозначим $x_*=x(\lambda^*)$. Тогда выполняется следующее соотношение:

$$\begin{gathered}f(x_\delta(\lambda))+\lambda g(x_\delta(\lambda))\leqslant f(x(\lambda))+\lambda g(x(\lambda))+\delta=\varphi(\lambda)+\delta\leqslant\\ \leqslant\varphi(\lambda^*)+\delta=f(x_*)+\lambda^*\underbrace{g(x_*)}_{\leqslant 0}+\delta\leqslant f(x_*)+\delta\end{gathered}$$

В силу критерия остановки алгоритма 15, получаем следующую оценку:

$$f(x_\delta(\lambda))-f(x_*)\leqslant-\lambda g(x_\delta(\lambda))+\delta\leqslant\delta+\varepsilon,$$

что соответствует первому неравенству леммы. Неравенство $g(x_\delta(\lambda))\leqslant\frac{\varepsilon}{\lambda}$ вытекает из критерия остановки алгоритма 15 для точки λ.

4.4.5 Оценка скорости сходимости в случае, когда $f(x)$ и $g(x)$ удовлетворяют условию Липшица Прежде чем оценить сложность рассматриваемого алгоритма, оценим сложность решения вспомогательной многомерной задачи. Точнее говоря, на каждой итерации алгоритма 15 возникает вспомогательная задача

$$x_\delta(\lambda) = \arg\min_{x \in Q} \left\{ f(x) + \lambda g(x) \right\},$$

необходимо решить с точностью δ по функции на каждой итерации, т.е. (4.43). Однако гарантировать выполнения критерия пункта 9 алгоритма 15 при нахождении $x(\lambda)$ с точностью δ по функции в общем случае невозможно. Поэтому для обоснования выполнимости критерия остановки метода будем требовать также и сходимость по аргументу вспомогательных подзадач:

$$||x_\delta(\lambda) - x(\lambda)||_2 \leqslant \delta.$$

Если знак $g(x_\delta(\lambda))$ совпадает со знаком $\varphi'(\lambda) = g(x(\lambda))$, то после соответствующей итерации искомое λ_* останется в промежутке локализации. Если же знаки $g(x_\delta(\lambda))$ и $g(x(\lambda))$ различны, то

$$|g(x_\delta(\lambda))| \leqslant |g(x_\delta(\lambda)) - g(x(\lambda))| \leqslant M_g \|x_\delta(\lambda) - x(\lambda)\|_2 \leqslant M_g \delta,$$

откуда для всякого $\lambda \leqslant \lambda_{max}$

$$\lambda |g(x_\delta(\lambda))| \leqslant \lambda_{max} M_g \delta.$$

Это означает, что при $\delta \leqslant \frac{\varepsilon}{\lambda_{max} M_g} = O(\varepsilon)$ можно гарантированно утверждать, что либо выполнится критерий остановки алгоритма 4, либо $g(x_\delta(\lambda))$ и $g(x(\lambda)) = \varphi'(\lambda)$ будут иметь одинаковый знак и искомое λ_* не выпадет из промежутка локализации для λ. Поэтому вполне подходящей будет выбор точности решения вспомогательных задач $\delta \leqslant \frac{\varepsilon}{\lambda_{max} M_g} = O(\varepsilon)$, что мы будем полагать далее.

При этом функция $F_\lambda(x) = f(x) + \lambda g(x)$ сильно выпукла и удовлетворяет условию Липшица для любого фиксированного λ ввиду липшицевости $f(x)$ и $g(x)$. Для решения вспомогательной задачи минимизации функции $F_\lambda(x)$ будем использовать метод проекции субградиента [72]. Хорошо известно, что после k итераций данного метода верна оценка:

$$F_\lambda(x^k) - F_\lambda(x_*) \leqslant \frac{2M_{F_\lambda}^2}{k\mu_f},$$

где $M_{F_\lambda} = M_f + \lambda M_g$. В силу сильной выпуклости функции $F_\lambda(x)$ справедливо следующее неравенство

$$F_\lambda(x) \geqslant F_\lambda(x_*) + \langle \nabla F_\lambda(x_*), x - x_* \rangle + \frac{\mu_f}{2}||x - x_*||_2^2 \geqslant F_\lambda(x_*) + \frac{\mu_f}{2}||x - x_*||_2^2.$$

Таким образом, получаем

$$||x - x_*||_2^2 \leqslant \frac{2}{\mu_f}\left(F_\lambda(x) - F_\lambda(x_*)\right).$$

Тогда для $x = x^k$ можно получить следующую оценку:

$$||x^k - x_*||_2^2 \leqslant \frac{4M_{F_\lambda}^2}{k\mu_f^2}.$$

Отсюда получаем, что количество итераций метода проекции субградиента для решения вспомогательной задачи с точностью δ по аргументу не превосходит

$$k = \frac{4M_{F_\lambda}^2}{\delta^2\mu_f^2}. \tag{4.44}$$

Теперь перейдем к анализу скорости сходимости алгоритма 15.

Отметим, что основной идеей метода дихотомии является уменьшение отрезка локализации решения в два раза за итерацию метода. При этом решение задачи всегда содержится в текущем отрезки локализации. Заметим, что при решении рассматриваемой двойственной задачи (4.40), параметр сильной выпуклости вспомогательной задачи зависит от λ. В этом случае, при приближении этого значения к нулю, параметр сильной выпуклости вспомогательной задачи будет также стремиться к нулю. В связи с этим, при анализе сходимости метода, будут рассмотрены два случая.

1. Предположим, что на каждой итерации метода дихотомии отбрасывается правая половина отрезка, то есть отрезок локализации λ ограничем снизу нулем. Тогда спустя N итераций отрезок локализации решения будет соответственно $\left[\lambda_{min}^N, \lambda_{max}^N\right] = \left[0, \frac{\lambda_{max}}{2^N}\right]$. Это позволяет получить следующую оценку (в силу критерия остановки алгоритма 15):

$$\lambda^N \left|\varphi'(\lambda^N)\right| \leqslant \lambda^N C_g \leqslant \varepsilon,$$

где $|g(x)| \leqslant C_g \quad \forall\, x \in Q$ для некоторой постоянной $C_g > 0$.

Отсюда при $\lambda^N \leqslant \frac{\lambda_{max}}{2^N}$ получаем, что критерий остановки метода дихотомии будет выполнен не позднее, чем спустя

$$N = \log_2 \frac{C_g \lambda_{max}}{\varepsilon}$$

итераций. Заметим, что в силу оценки (4.44) трудоемкость каждой вспомогательной задачи, возникающей в ходе работы алгоритма, можно оценить следующим образом:

$$k = \frac{4(M_f + \lambda_{max} M_g)^2}{\delta^2 \mu_f^2}$$

$$\log_2 \frac{2C_g \lambda_{max}}{\varepsilon} O\left(\frac{4(M_f + \lambda_{max} M_g)^2}{\delta^2 \mu_f^2}\right),$$

т.е. ввиду $\delta = O(\varepsilon)$ $O\left(\frac{1}{\varepsilon^2} \log_2 \frac{1}{\varepsilon}\right)$.

2. Теперь рассмотрим второй случай, когда на одной из итераций дихотомии была отброшена часть $\lambda \in [0; \frac{\lambda_{max}}{2^p}]$. Тогда для некоторого $p \leqslant N$ определим $\lambda^1 = \frac{\lambda_{max}}{2^p}$ и $\lambda^2 = \frac{\lambda_{max}}{2^{p-1}}$. Соответственно, $\varphi'(\lambda^1)$ и $\varphi'(\lambda^2)$ противоположны по знаку и $\lambda^* \in \left[\lambda^1, \lambda^2\right]$.

Из теоремы 4.4.3 вытекает следующая оценка:

$$|\varphi'(\lambda_2) - \varphi'(\lambda_1)| \leqslant \frac{M_g^2}{\mu_f} |\lambda_2 - \lambda_1| \quad \forall\, \lambda_1,\, \lambda_2 \ \in [\lambda_{min}, \lambda_{max}].$$

Тогда

$$\begin{aligned} |\lambda_2 \varphi'(\lambda_2) - \lambda_1 \varphi'(\lambda_1)| &\leqslant |\lambda_2 \varphi'(\lambda_2) - \lambda_2 \varphi'(\lambda_1)| + |\lambda_2 \varphi'(\lambda_1) - \lambda_1 \varphi'(\lambda_1)| \leqslant \\ &\leqslant \lambda_{max} |\varphi'(\lambda_2) - \varphi'(\lambda_1)| + |\varphi'(\lambda_1)| \cdot |\lambda_2 - \lambda_1| \leqslant \\ &\leqslant \left(\frac{M_g^2 \lambda_{max}}{\mu_f} + |\varphi'(\lambda_1)|\right) |\lambda_2 - \lambda_1| = C\, |\lambda_2 - \lambda_1|, \end{aligned}$$

где $C = \frac{M_g^2 \lambda_{max}}{\mu_f} + C_g$. Подставляя $\lambda_2 = \lambda^N$, $\lambda_1 = \lambda^*$ и учитывая, что в силу работы метода дихотомии $\lambda^N,\, \lambda^* \in \left[\lambda_{min}^N, \lambda_{max}^N\right]$, получаем

$$\lambda^N \left|\varphi'(\lambda^N)\right| \leqslant C \left|\lambda^N - \lambda^*\right| \leqslant C \left|\lambda_{max}^N - \lambda_{min}^N\right|.$$

В этом случае критерий остановки будет выполнен после

$$\lambda_{max}^N - \lambda_{min}^N \ = \frac{\lambda_{max}}{2^N} \leqslant \frac{\varepsilon}{C},$$

т.е. $N = \log_2 \dfrac{C\lambda_{max}}{\varepsilon}$ итераций.

При этом константу сильной выпуклости вспомогательной задачи можно оценить как $\mu_F = \mu \dfrac{\lambda_{max}}{2^{N-1}}$. Тогда справедливы следующие оценки:

$$2^N \leqslant \frac{C\lambda_{max}}{\varepsilon}, \quad 2^N \geqslant \frac{C\lambda_{max}}{2\varepsilon}.$$

Отсюда сложность каждой вспомогательной задачи можно оценить как

$$k = \frac{4(M_f + M_g)^2}{\delta^2 \mu_f^2}, \quad p = 0, ..., N.$$

Поэтому получаем оценку сложности

$$\log_2 \frac{C_g \lambda_{max}}{\varepsilon} O\left(\frac{4\left(M_f + \lambda_{max} M_g\right)^2}{\delta^2 \mu_f^2}\right), \quad \text{то есть} \quad O\left(\frac{1}{\varepsilon^2} \log_2 \frac{1}{\varepsilon}\right).$$

4.4.6 Оценка скорости сходимости для случая гладких функций Предположим, что функции f и g гладкие, т.е. существуют такие константы L_f и L_g, что

$$||\nabla f(y) - \nabla f(x)||_2 \leqslant L_f ||y - x||_2, \ \ \forall x, y \in Q,$$

$$||\nabla g(y) - \nabla g(x)||_2 \leqslant L_g ||y - x||_2, \ \ \forall x, y \in Q.$$

Прежде чем перейти к дальнейшим оценкам скорости сходимости, заметим, что прямая задача решается на компакте Q, отсюда получаем следующую оценку:

$$d \overset{def}{=} \max_{x,y \in Q} ||x - y||_2. \tag{4.45}$$

На каждой итерации алгоритма 15 возникает вспомогательная задача

$$x_\delta(\lambda) = \arg\min_{x \in Q} \left\{ f(x) + \lambda g(x) \right\}.$$

При этом в данном случае функция $F_\lambda(x) = f(x) + \lambda g(x)$ будет сильно выпуклой и гладкой для любого фиксированного λ в силу свойств функций $f(x)$ и $g(x)$.

Вспомогательную задачу необходимо решать с точностью $\delta = O(\varepsilon)$ по аргументу на каждой итерации ($||x_\delta(\lambda) - x(\lambda)||_2 \leqslant \delta$).

Для решения вспомогательной задачи минимизации функции $F_\lambda(x)$ будем использовать ускоренный градиентный метод ([83], п. 3.7.1), для скорости сходимости которого известна оценка:

$$||x^k(\lambda) - x(\lambda)||_2^2 \leqslant \frac{L_{F_\lambda} + \mu_f}{2}||x^0(\lambda) - x(\lambda)||_2^2 \left(1 - \sqrt{\frac{\mu_f}{L_{F_\lambda}}}\right)^k,$$

где $L_{F_\lambda} = L_f + \lambda L_g$. Тогда точность δ решения вспомогательной задачи по аргументу достигается после

$$k = \sqrt{\frac{L_{F_\lambda}}{\mu_f}} \log_2 \left(\frac{d}{\delta}\right),$$

итераций, где d определяется в (4.45). Далее оценка трудоёмкости метода получается аналогично пункту 4.4.5 и при $\delta = O(\varepsilon)$ принимает вид $O\left(\log_2^2 \frac{1}{\varepsilon}\right)$.

4.4.7 Оценка для задач композитной оптимизации

Выделим также важный частный случай. Пусть f имеет липшицев градиент с константой L_f

$$||\nabla f(x) - \nabla f(y)||_2 \leqslant L_f||x - y||_2 \quad \forall\, x,\, y \in Q$$

и $g(x)$ так называемая простая функция, т.е. негладкая выпуклая функция простой структуры. Последнее означает, что множества Лебега

$$\Lambda_y = \{x \in Q : g(x) < y\}$$

имеют простую структуру. Например, к таким проблемам можно отнести проблему известную задачу:

$$\frac{1}{2}||Ax - b||_2^2 + \lambda||x||_1 \to \min_{x \in \mathbb{R}^n},$$

где A — матрица размерности $(m \times n)$, $b \in \mathbb{R}^m$, λ — параметр регуляризации, а $||\cdot||_1$ обозначает стандартную l_1-норму. Тогда мы можем использовать следующую процедуру градиентного типа

$$x^{k+1} = \arg\min_{x \in Q} \left\{\langle \nabla f(x^k), x - x^k\rangle + \lambda g(x) + \frac{L_f}{2}||x - x^k||_2^2\right\}. \qquad (4.46)$$

Для метода (4.46) мы можем получить $||x - x(\delta)||_2 \leqslant \varepsilon$ после

$$\frac{L_f}{\lambda\mu}\log_2\frac{1}{\delta}$$

итераций метода (4.46). В таком случае общая трудоемкость (количество обращений к (суб)градиенту $f(x)$ и $g(x)$) алгоритма 15 будет иметь вид:

$$O\left(\log_2^2\frac{1}{\varepsilon}\right). \tag{4.47}$$

Скорость сходимости аналогична в случае, когда $g(x)$ — гладкая выпуклая функция простой структуры (см. (4.4.7)). Пусть g имеет липшицев градиент с константой L_g

$$||\nabla g(x) - \nabla g(y)||_2 \leqslant L_g||x - y||_2 \ \forall x, y \in Q$$

и f негладкая выпуклая функция. Тогда мы можем использовать следующую процедуру градиентного типа

$$x^{k+1} = \arg\min_{x\in Q}\left\{\langle\lambda\nabla g(x^k), x - x^k\rangle + f(x) + \frac{\lambda L_g}{2}||x - x^k||_2^2\right\}. \tag{4.48}$$

Таким образом, для метода (4.48) мы можем получить качество решения $||x - x(\delta)||_2 \leqslant \varepsilon$ после

$$\frac{L_g}{\mu_f}\log_2\frac{1}{\delta}$$

итераций метода (4.48). В этом случае оценка (4.47) для алгоритма 15 также будет иметь вид $O\left(\log_2^2\frac{1}{\varepsilon}\right)$.

Замечание 4.4.5. Заметим, что вместо обычного процедуры градиентного типа (4.48) можно использовать технику рестартов хорошо известного быстрого градиентного метода Ю. Е. Нестерова. Этот подход применим и для композитной оптимизации (см. [18]). Это не изменит оценки трудоемкости (4.47), однако позволит уменьшить константы, связанные с числами обусловленности вспомогательных подзадач: вместо $\frac{L_f}{\mu}$ и $\frac{\lambda L_g}{\mu_f}$ можно использовать квадратные корни из этих величин. Однако для запуска ускоренного метода знать параметр сильной выпуклости μ или его нижнюю оценку. Неускоренный градиентный метод можно запускать и без знания этого параметра.

4.4.8 Численные эксперименты. Сравнение с универсальным методом для седловых задач Рассматриваемую задачу с функционалом ограничения $g(x) = \max\limits_{1\leqslant p\leqslant m}\{g_p(x)\}$ можно свести к седловой задаче следующего вида $\min\limits_{x\in Q}\max\limits_{\lambda} F(x,\lambda)$, где $F(x,\lambda) = f(x) + \sum\limits_{p=1}^{m}\lambda_p g_p(x)$, $\overrightarrow{\lambda} = (\lambda_1,\lambda_2,\ldots,\lambda_m)^T \in \mathbb{R}^m_+$. Отметим, что в случае гладких $g_p(x)$ ($p = \overline{1,m}$) замена системы ограничений на одно $g(x) = \max\limits_{1\leqslant p\leqslant m}\{g_p(x)\}$ может привести к понижению уровня гладкости седловой задачи, что приведет к худшей оценке скорости сходимости. Указанного типа седловую задачу можно решать с использованием рассмотренного выше в разделе 2 универсального метода для соответствующего вариационного неравенства: $\langle G(x_*, \overrightarrow{\lambda}_*), (x_*, \overrightarrow{\lambda}_*) - (x, \overrightarrow{\lambda})\rangle \leqslant 0 \ \ \forall (x, \overrightarrow{\lambda}) \in B \subset \mathbb{R}^{n+m}$, где $B = \left\{(x, \overrightarrow{\lambda}) \,|\, \sum\limits_{k=1}^{n} x_k^2 + \sum\limits_{p=1}^{m}\lambda_p^2 \leqslant 1\right\}$, и оператор G имеет следующий вид

$$G(x,\lambda) = \begin{pmatrix} \nabla f(x) + \sum\limits_{p=1}^{m}\lambda_p \nabla g_p(x), \\ (-g_1(x), -g_2(x), \ldots, -g_m(x))^T \end{pmatrix}.$$

Для сравнения предложенной методики с универсальным методом совместно с аспирантом МФТИ Мохаммадом Алкусой были проведены численные эксперименты. Выбрана выпуклая и липшицева (негладкая) целевая функция f — и функционал ограничения g — сильно выпуклый и имеет следующий вид: $g(x) = \max\limits_{1\leqslant i\leqslant m}\{g_i(x)\}$, где

$$g(x) = \max_{i\in[m]}\{g_i(x) = \alpha_{i1}x_1^2 + \alpha_{i2}x_2^2 + \ldots + \alpha_{in}x_n^2 - 10\}, \tag{4.49}$$

Коэффициенты α_i взяты с равномерным распределением по полуоткрытому интервалу $[0,1)$. Рассмотрим эксперименты с тремя различными целевыми функционалами.

Пример 4.4.6. Аналог задачи Ферма–Торричелли–Штейнера с ограничениями. Этот пример связан с аналогом известной задачи Ферма–Торричелли–Штейнера с некоторыми функциональными ограничениями. Для заданного набора $\{P_k = (p_{1k}, p_{2k}, \ldots, p_{nk});\, 1 \leqslant k \leqslant r\}$ r точек в n-мерном евклидовом пространстве $\mathbb{R}^n$ нам нужно решить оптимизационную задачу

$$f(x) \to \min_{x\in Q,\, g(x)\leqslant 0}, \tag{4.50}$$

где целевой функционал f имеет вид

$$f(x) := \frac{1}{r}\sum_{k=1}^{r}\sqrt{(x_1-p_{1k})^2+\ldots+(x_n-p_{nk})^2} = \frac{1}{r}\sum_{k=1}^{r}\|x-P_k\|_2\,,$$

где координаты точек P_k для всех k, $1 \leqslant k \leqslant r$ взяты с нормальным (гауссовским) распределением в евклидовом единичном шаре с центром в нуле.

Пример 4.4.7. Аналог задачи о наименьшем покрывающем шаре с ограничениями. Для заданного набора $\{P_k = (p_{1k}, p_{2k}, \ldots, p_{nk});\, 1 \leqslant k \leqslant r\}$ r точек в n-мерном евклидовом пространстве $\mathbb{R}^n$ нам нужно решить оптимизационную задачу (4.50), где целевой функционал f имеет вид

$$f(x) := \max_{1\leqslant k\leqslant r}\left\{\sqrt{(x_1-p_{1k})^2+\ldots+(x_n-p_{nk})^2}\right\},$$

где координаты точек P_k для всех $1 \leqslant k \leqslant r$ взяты из нормального (гауссовского) распределения.

Результаты работы алгоритмов 15 и 16 для примеров 4.4.6 и 4.4.7 представлены в таблице 4.5.

Отметим, что при реализации алгоритма 15 на каждой итерации (см. пункт 4 в алгоритме 15), для решения вспомогательной задачи мы использовали метод проекции субградиента[1)] с начальной точкой $x^0 = \left(\frac{1}{\sqrt{n}},\ldots,\frac{1}{\sqrt{n}}\right) \in Q \subset \mathbb{R}^n$.

Для алгоритма 15 выбираем $\lambda_{min} = 0$, $\lambda_{max} = \frac{f(\bar{x})}{-g(\bar{x})}$, где $\bar{x}$ — произвольная точка, такая, что $g(\bar{x}) < 0$.

[1)] Для этого алгоритма мы рассмотрим следующую схему [83]

$$x_{k+1} = \mathrm{Proj}_Q\left(x_k - h_k\nabla F(x_k)\right),$$

где F — μ-сильно выпукло и M-липшицево, $\nabla F(x_k)$ — некоторый субградиент F в точкеx_k. Размер шага $h_k = \frac{2}{\mu(k+1)}$. По этой стратегии мы имеем следующую скорость сходимости:

$$F\left(\sum_{k=1}^{N}\frac{2k}{N(N+1)}x_k\right) - F^* \leqslant \frac{2M^2}{\mu(N+1)}.$$

В экспериментах выбраны параметры $N = 1000$ для $\varepsilon = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$, $N = 5000$ для $\varepsilon = \frac{1}{10}, \frac{1}{16}$, $N = 10\,000$ для $\varepsilon = \frac{1}{32}, \frac{1}{64},, \frac{1}{100}$, $N = 100\,000$ для $\varepsilon = \frac{1}{128}, \frac{1}{256}, \frac{1}{1000}$ и $N = 1000\,000$ для $\varepsilon = \frac{1}{10\,000}$.

Для алгоритма 16 выбираем $L_0 = \frac{\|G(1,0,\ldots,0)-G(0,1,0,0,\ldots,0)\|}{\sqrt{2}}$ и начальную точку

$$(x^0, \overrightarrow{\lambda}^0) = \left(\frac{1}{\sqrt{m+n}}, \frac{1}{\sqrt{m+n}} \ldots, \frac{1}{\sqrt{m+n}}\right) \in \mathbb{R}^{n+m}.$$

Таблица 4.5. Результаты работы алгоритма 15 и 16.

$\frac{1}{\varepsilon}$	Пример 4.4.6. $m = n = 500, r = 10.$							
	Алгоритм 15				Алгоритм 16			
	N	Время, с	f^{best}	g^{out}	N	Время, с	f^{best}	g^{out}
2	6	261	21.843	-9.446	5	8	22.22590388	-9.99999780
4	7	303	21.841	-9.440	6	10	22.22591070	-9.99999858
8	8	3375	21.841	-9.438	7	12	22.22591403	-9.99999878
16	—	—	—	—	8	13	22.22591568	-9.99999884
32	—	—	—	—	9	16	22.22591650	-9.99999885
64	—	—	—	—	10	18	22.22591691	-9.99999886
128	—	—	—	—	11	20	22.22591712	-9.99999886
256	—	—	—	—	12	21	22.22591722	-9.99999886
	Пример 4.4.7. $m = n = 500, r = 10.$							
2	6	96	22.968	-9.757	5	6	23.30771013	-9.99998759
4	7	112	22.959	-9.745	6	7	23.30772512	-9.99998839
8	8	1256	22.954	-9.738	7	8	23.30773245	-9.99998861
16	—	—	—	—	8	9	23.30773607	-9.99998868
32	—	—	—	—	9	11	23.30773787	-9.99998870
64	—	—	—	—	10	11	23.30773876	-9.99998871
128	—	—	—	—	11	13	23.30773921	-9.99998871
256	—	—	—	—	12	15	23.30773944	-9.99998871

Алгоритм 16 Универсальный метод для вариационных неравенств

Require: точность $\varepsilon > 0$, точка старта x^0, $L_0 > 0$, $R^2 = \frac{1}{2}\max\limits_{x \in Q}\|x - x^0\|^2$.

1: Задаем $N = 0$.

2: Задаем $i_N = 0$.

3: Задаем $L_{N+1} = 2^{i_N - 1} L_N$.

4: Вычисляем

$$y^{N+1} = \arg\min_{x \in Q} \left\{ \langle G(x^N), x - x^N \rangle + \frac{L_{N+1}}{2}\|x^N - x\|^2 \right\}.$$

5: Вычисляем

$$x^{N+1} = \arg\min_{x \in Q} \left\{ \langle G(y^{N+1}), x - x^N \rangle + \frac{L_{N+1}}{2}\|x^N - x\|^2 \right\}.$$

6: $i_N := i_N + 1$.

$$\langle G(y^{N+1}) - G(x^N), y^{N+1} - x^{N+1} \rangle \leqslant \frac{L_{N+1}}{2}\|x^N - y^{N+1}\|^2 +$$

$$+ \frac{L_{N+1}}{2}\|x^{N+1} - y^{N+1}\|^2 + \frac{\varepsilon}{2}.$$

7: Задаем $N = N + 1$. $\quad \sum_{N=0}^{k-1} \frac{1}{L_{N+1}} \geqslant \frac{2R^2}{\varepsilon}$

Ensure: $\widehat{y} = \frac{1}{\sum_{N=0}^{k-1} \frac{1}{L_{N+1}}} \sum_{N=0}^{k-1} \frac{y^{N+1}}{L_{N+1}}$.

Для сравнения алгоритмов 15 и 16 были проведены численные эксперименты, для которых целевая функция f — выпуклая и липшицева (вообще говоря, негладкая) и функционал ограничения g — сильно выпуклый и имеет следующий вид:

$$g(x) = \max_{1 \leqslant i \leqslant m} \{g_i(x)\}, \tag{4.51}$$

где

$$g_i(x) = \alpha_{i1}x_1^2 + \alpha_{i2}x_2^2 + \ldots + \alpha_{in}x_n^2 - 10; \quad i = \overline{1, m}. \tag{4.52}$$

Коэффициенты $\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in}$, при всех $i = \overline{1, m}$, взяты с равномерным распределением по полуоткрытому интервалу $[0, 1)$. Рассмотрим несколько разных примеров с разными негладкими выпуклыми целевыми функциями.

Пример 4.4.8.

$$f(x) = \exp^{\|x\|_2}.$$

Пример 4.4.9.

$$f(x) = \frac{1}{r}\sum_{i=1}^{r} |a_{i1}\,x_1 + a_{i2}\,x_2 + \ldots + a_{in}\,x_n|,$$

где коэффициенты $a_{i1}, a_{i2}, \ldots, a_{in}$ для всех $i = \overline{1,r}$ взяты из равномерного распределения по полуоткрытому интервалу $[0,1)$ с нормализацией для получения $M_f = 1$, которая представляет собой константу Липшица целевой функции f.

Выберем допустимое множество

$$Q = \{x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n\,;\, x_1^2 + x_2^2 + ... + x_n^2 \leqslant 1\}.$$

Обратите внимание, что для примеров 4.4.8 и 4.4.9 с функциональными ограничениями (4.51)–(4.52) точное решение рассматриваемой задачи (4.35) достигается в точке $x_* = \mathbf{0} \in \mathbb{R}^n$.

Для примеров 4.4.8 и 4.4.9 мы запускаем алгоритмы 15 и 16 для $n = 200$. Результаты работы алгоритмов 15 и 16 представлены в таблице 4.6 ниже.

Эти результаты демонстрируют сравнение времени работы (в секундах) алгоритмов 15 и 16 с различными значениями ε, а также качества решения, которое создается этими алгоритмами с учетом целевой функции f и функциональных ограничений g, где мы вычисляем значения этих функций на выходе алгоритмов. Для алгоритма 15 устанавливаем: $\lambda^{\text{out}} := \lambda^N$, $f^{\text{best}} := f\left(x_\delta(\lambda^{\text{out}})\right)$ и $g^{\text{out}} := g\left(x_\delta(\lambda^{\text{out}})\right)$. Для алгоритма 16 устанавливаем: $f^{\text{best}} := f\left(\widehat{y}\right)$ и $g^{\text{out}} := g\left(\widehat{y}\right)$, где $\widehat{y}$ — выходная точка алгоритма 16.

Когда время выполнения алгоритма 16 очень велико, мы останавливаем их работу, что соответствует символу «—» в таблице 4.6.

Из таблицы 4.6, во-первых, мы можем видеть, что, поскольку $g^{\text{out}} := g\left(x_\delta(\lambda^N)\right) < 0$, мы имеем гарантию достижения критериев остановки алгоритма 15. Кроме того, мы можем видеть, что в общем случае для примеров 4.4.8 и 4.4.9 предложенный алгоритм 15 работает лучше, чем алгоритм 16 относительно времени выполнения (обратите внимание, что разница между временем выполнения каждого алгоритма

Таблица 4.6. Результаты работы алгоритмов 15 и 16 для примеров 4.4.8 и 4.4.9.

$\frac{1}{\varepsilon}$	Пример 4.4.8. $n = 200, m = 10$.					
	Алгоритм 15			Алгоритм 16		
	Время, с	f^{best}	g^{out}	Время, с	f^{best}	g^{out}
2	0.580	1.0009998401	-9.9999990218	0.032	1.0348541235	-9.9993775789
4	0.850	1.0009997788	-9.9999990415	0.410	1.0317138799	-9.9994833297
8	1.006	1.0009997067	-9.9999990742	13.367	1.0066641652	-9.9999765845
10	4.982	1.0001999943	-9.9999999606	16.795	1.0064118458	-9.9999783270
16	5.940	1.0001999920	-9.9999999613	56.872	1.0041653353	-9.9999908337
32	13.211	1.0000999980	-9.9999999903	319.763	1.0009455212	-9.9999995251
64	15.282	1.0000999973	-9.9999999906	1580.662	1.0004111180	-9.9999999101
100	17.273	1.0000999964	-9.9999999910	2278.351	1.0003496445	-9.9999999350
128	209.708	1.0000099999	-9.9999999999	6423.017	1.0002507367	-9.9999999665
256	220.554	1.0000099999	-9.9999999999	—	—	—
1000	255.474	1.0000099999	-9.9999999999	—	—	—
10 000	3230.091	1.0000009999	-9.9999999999	—	—	—
	Пример 4.4.9. $n = 200, m = 10, r = 5$.					
2	0.456	0.0008938886	-9.9999994423	0.031253	0.0263978538	-9.9995346497
4	0.931	0.0008941490	-9.9999994499	0.031254	0.0263978538	-9.9995346497
8	1.400	0.0008942146	-9.9999994537	8.411448	0.0116838336	-9.9999086899
10	8.747	0.0001789892	-9.9999999781	10.60636	0.0113375052	-9.9999142175
16	8.709	0.0001789892	-9.9999999781	66.82032	0.0049482809	-9.9999835825
32	21.850	8.950399e-05	-9.9999999945	238.4323	0.0022850773	-9.9999964943
64	26.193	8.950409e-05	-9.9999999945	939.4467	0.0008160358	-9.9999995497
100	30.373	8.950411e-05	-9.9999999945	2708.537	0.0004966824	-9.9999998329
128	154.261	1.790225e-05	-9.9999999997	—	—	—
256	176.288	1.790225e-05	-9.9999999997	—	—	—
1000	220.823	1.790225e-05	-9.9999999997	—	—	—
10 000	3087.790	1.790257e-06	-9.9999999999	—	—	—

очень велика, когда желаемая точность решения увеличивается) и относительно качества решения для целевой функции и функциональных ограничений (напомним, что для примера 4.4.8 имеем $x_* = \mathbf{0}$ и $f^* = 1, g(x_*) = -10$. Также для примера 4.4.9 имеем $x_* = \mathbf{0}$ и $f^* = 0, g(x_*) = -10$).

4.5 Аналог дихотомии для двумерной минимизации на квадрате и его приложения к задачам выпуклого программирования с двумя функционалами ограничений

Настоящий раздел посвящён задаче следующего вида

Пусть рассматривается задача

$$f(x) \to \min,\ x \in Q,\ g_1(x) \leqslant 0,\ g_2(x) \leqslant 0, \tag{4.53}$$

где f, g_1 и g_2 — выпуклые функционалы, Q — выпуклый компакт в $\mathbb{R}^n$. Допустим также, что f — μ-сильно выпукла относительно стандартной евклидовой нормы и $\forall i = \overline{1,2}$

$$|g_i(x) - g_i(y)| \leqslant M_i \|x - y\|_2$$

при всяких $x, y \in Q$ для фиксированных $\mu_{1,2} > 0$. Тогда двойственная задача к 4.53 имеет вид

$$\varphi(\lambda_1, \lambda_2) = \min_{x \in Q} \{f(x) + \lambda_1 g_1(x) + \lambda_2 g_2(x)\} \to \max_{\lambda_1 + \lambda_2 \leqslant \Omega_\lambda}, \tag{4.54}$$

где Ω_λ определяется условием Слейтера.

Функция φ зависит от двух двойственных переменных $\lambda = (\lambda_1, \lambda_2)$ и удовлетворяет как условию Липшица, так и имеет липшицев градиент:

$$|\varphi(\lambda) - \varphi(\nu)| \leqslant C \|\lambda - \nu\|_2,$$

а также

$$\|\nabla\varphi(\lambda) - \nabla\varphi(\nu)\|_2 \leqslant \frac{M^2}{\mu} \|\lambda - \nu\|_2$$

для произвольных $\lambda, \nu \in \mathbb{R}^2$, где $C = \max_{x \in Q} \{g_1(x), g_2(x)\}$ и $M = \sqrt{M_1^2 + M_2^2}$.

Если для прямой задачи выполняется условие Слейтера, то можно получить ограничение следующего вида для точки-решения двойственной задачи:

$$\lambda_1^* + \lambda_2^* \leqslant A.$$

Таким образом, решение задачи сводится к оптимизации функции φ двух переменных на прямоугольном треугольнике

$$\Omega_\lambda = \Lambda_A = \{\lambda \in \mathbb{R}_+^2 | \lambda_1 + \lambda_2 \leqslant A\},$$

катеты которого равны A и лежат на осях координат. Можно рассматривать 2-мерную двойственную задачу на квадрате, содержащем указанный прямоугольный треугольник.

В качестве аналога дихотомии для необходимой двумерной минимизации будем использовать предложенный Ю. Е. Нестеровым неполноградиентный метод минимизации липшицевой выпуклой липшицевой функции двух переменных на квадрате [44]. Особенность этого метода — сведение решаемой задачи минимизации к постепенному сужению области определения за счёт учёта направления (суб)градиентов целевой функции вблизи точек-решений вспомогательных одномерных задач. Иными словами, происходит комбинирование быстро работающего метода одномерной минимизации и с учётом геометрической структуры области определения задачи выяснение направления субградиентов в текущих точках на каждой итерации. Важно, что на каждой итерации метода два раза находится направление градиента (а не сам градиент), что существенно удешевляет стоимость итерации по сравнению с полноградиентным методом. При этом для выпуклых и гладких функций сохраняется линейная скорость сходимости метода. Подходы предыдущего пункта работы применимы и для задач с двумя ограничениями. Однако для этого необходимо предложить аналог метода дихотомии, причем для возмущенного значения градиента. Этой цели и посвящен настоящий раздел работы.

4.5.1 Описание метода Дан квадрат $\Pi \subset \mathbb{R}^2$ со стороной R, задана функция $f : \Pi \to \mathbb{R}$. Опишем итерацию предлагаемого метода. Сформулируем его как для гладких, так и негладких целевых функционалов, хотя метод может расходиться на негладкой функции (см.

пример далее). При этом отметим, что для негладких выпуклых функционалов вместо обычного градиента используется *субградиент* $\upsilon \in \mathbb{R}^2$ определяемый неравенством

$$f(x) - f(x_0) \geq \langle \upsilon, x - x_0 \rangle,$$

справедливым для всех x из области определения.

1. Через центр имеющегося квадрата проводится горизонтальная прямая. На отрезке, высекаемом из квадрата этой прямой, с точностью δ по аргументу решается задача одномерной оптимизации. Далее при реализации для решения одномерной задачи мы используем метод золотого сечения.

2. В найденной точке x_δ вычисляется вектор (суб)градиента $\nabla f(x_\delta)$ и определяется, в сторону какого из двух прямоугольников направлен вектор $\nabla f(x_\delta)$ и этот прямоугольник исключается из рассмотрения. При этом важно, что достаточно знать именно направление, а не точное значение (суб)градиента.

3. Через центр оставшегося прямоугольника проводится вертикальная прямая, на отрезке, высекаемом этой прямой в прямоугольнике, также с точностью δ по аргументу решается задача одномерной оптимизации. В найденной точке вычисляется вектор (суб)градиента функции и определяется, в сторону какого из двух квадратов он направлен. Этот квадрат исключается из рассмотрения.

Если вектор (суб)градиента в точке-приближении решения вспомогательной одномерной задачи нулевой, то процесс можно остановить и выдать указанную точку (она и даст точное решение).

Также возможно, что (суб)градиент в данной точке будет направлен вдоль отрезка, на котором решается вспомогательная одномерная задача и неясно, какую часть оставшейся фигуры отсекать. В таком случае условимся отсекать любую из частей на выбор.

4.5.2 Обоснование оценки скорости сходимости

Легко понять, что на каждой итерации метода Ю. Е. Нестерова линейные размеры квадрата уменьшаются вдвое и в какой-то момент оставшийся квадрат будет настолько мал, что в силу условия Липшица значения целевого функционала в его точка будут достаточно близкими друг к другу и можно выбирать любую точку. Мы докажем, что при подходящем выборе точности решения вспомогательных одномерных задач

можно гарантировать, что значения целевой функции в точках оставшегося квадрата будут достаточно близки к оптимальному. Следующий результат получен нами совместно с Дмитрием Аркадьевичем Пасечнюком в [44].

Теорема 4.5.1. *Пусть на некотором квадрате $\Pi \subset \mathbb{R}^2$ со стороной R задана выпуклая дифференцируемая функция $f : \Pi \to \mathbb{R}$, для которой при любых $x, y \in \Pi$*

1) $|f(x) - f(y)| \leqslant M\|x - y\|_2$,

2) $\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|_2$.

Тогда для всякого $\varepsilon > 0$ после

$$n = \log_2 \frac{2MR\sqrt{2}}{\varepsilon}$$

шагов метода Ю. Е. Нестерова (M выбрана так, чтобы n было натуральным) при условии, что вспомогательные одномерные задачи решаются с точностью по аргументу

$$\delta = \frac{\varepsilon}{2LR\left(\sqrt{2} + \sqrt{5}\right)\left(1 - \dfrac{\varepsilon}{MR\sqrt{2}}\right)}, \tag{4.55}$$

получим квадрат $K_n \subset \Pi = K_0$ такой, что

$$f(x) - \min_{x \in \Pi} f(x) \leqslant \varepsilon \quad \forall x \in K_n.$$

Замечание 4.5.2. Из доказанного результата по стандартной схеме рассуждений можно доказать и сходимость метода по аргументу в классе сильно выпуклых целевых функционалов.

Замечание 4.5.3. Важно отметить, что схема доказательства теоремы 4.5.1 (см. [44]) позволяет учитывать не только погрешность решения вспомогательных одномерных задач, но и погрешность вычисления градиента в точках x_δ (см. пп. 3 и 4 доказательства теоремы 4.5.1 в [44]). Точнее говоря, в точках x_δ вектор градиента функции может вычисляться с точностью Δ: если $\nabla f(x_\delta)$ — точная величина градиента, то нам доступен некоторый вектор $v(x_\delta)$, для которого $\|v(x_\delta) - \nabla f(x_\delta)\|_2 \leqslant \Delta$.

В том случае, если вектор $v(x_\delta)$ направлен в ту же часть делимого квадрата, что и вектор $\nabla f(x_\delta)$, все рассуждения п. 3 доказательства теоремы 4.5.1 остаются неизменными, так как исключается та же часть квадрата, что и при $\Delta = 0$. Если же это не так, то либо вектор $\nabla f(x_\delta)$ направлен в ту часть, где лежит x_* – точное решение задачи, а вектор $v(x_\delta)$ – в иную (очевидно, что в результате будет исключена часть, которой не принадлежит x_*, так что рассуждения в п. 3 также остаются справедливыми), либо же вектор $v(x_\delta)$ в ту часть, где лежит x_*, а $\nabla f(x_\delta)$ – в иную.

В таком случае имеем треугольник, образованный векторами $\nabla f(x_\delta) - \nabla f(x_0)$ и $v(x_\delta) - \nabla f(x_\delta)$. Из неравенства треугольника следует, что если $\|\nabla f(x_\delta) - \nabla f(x_0)\|_2 \leqslant L\delta$ и $\|v(x_\delta) - \nabla f(x_\delta)\|_2 \leqslant \Delta$, то

$$\|v(x_\delta) - \nabla f(x_0)\|_2 \leqslant L\delta + \Delta.$$

Векторы $v(x_\delta)$ и $\nabla f(x_0)$ образуют тупоугольный треугольник, причём длина его наибольшей стороны $\|v(x_\delta) - \nabla f(x_0)\|_2 \leqslant L\delta + \Delta$. Следовательно, $\|v(x_\delta)\|_2 \leqslant L\delta + \Delta$. Из треугольника, образуемого векторами $v(x_\delta)$ и $\nabla f(x_\delta)$ ввиду неравенства треугольника, имеем

$$\|\nabla f(x_\delta)\|_2 \leqslant \|v(x_\delta)\|_2 + \|v(x_\delta) - \nabla f(x_\delta)\|_2 \leqslant L\delta + 2\Delta.$$

Далее, следуя схеме рассуждений доказательства основной теоремы, получим

$$(L\delta + 2\Delta) \cdot R\left(\sqrt{2} + \sqrt{5}\right)\left(1 - \frac{\varepsilon}{MR\sqrt{2}}\right) \leqslant \frac{\varepsilon}{2},$$

откуда

$$2\Delta + L\delta \leqslant \frac{\varepsilon}{2R\left(\sqrt{2} + \sqrt{5}\right)\left(1 - \dfrac{\varepsilon}{MR\sqrt{2}}\right)}.$$

Таким образом, метод можно применить при наличии двух типов погрешностей: при решении вспомогательных одномерных задач оптимизации и при вычислении (направлений) градиентов. Эти два типа погрешностей аккумулируют итоговую погрешность. При увеличении одной из погрешности можно уменьшать значение другой так, чтобы сохранялось последнее неравенство, гарантирующее достижение приемлемого качества решения.

Отметим, что при этом можно построить такой пример функции, что для неё даже при небольшой погрешности вычисления градиента может не наблюдаться свойство сходимости метода по аргументу.

Пример 4.5.4. Рассмотрим на квадрате $[0;1]^2$ функцию

$$f(x_1, x_2) = x_1 - 0.001x_2,$$

минимум которой достигается в точке $(0;1)$. Градиент данной функции постоянен и равен $\nabla f(x) = (1; -0.001)$, причём вторая координата мала по модулю. Если при некотором Δ эта компонента будет найдена с погрешностью большей 0.001, то может оказаться, что она станет положительной и полученный вектор будет направлен в прямоугольник, содержащий точку точного решения задачи, в результате чего он будет исключен из рассмотрения. Из-за этого после удаления прямоугольника, содержащего точное решение задачи на первой итерации, метод не будет сходиться по аргументу.

Тем не менее, по функции будет достигнуто необходимое качество решения, поскольку ввиду малого коэффициента при x_2 слагаемое $0.001x_2$ не будет значительно влиять на значение функции.

Замечание 4.5.5. Функция, рассмотренная в предыдущем примере, обладает также тем свойством, что константа Липшица её градиента равна 0, что позволяет выбирать в качестве значения L любую сколь угодно малую величину. Поскольку $\delta \sim \frac{1}{L}$ (см. (4.55)), то в этом случае погрешность решения одномерной подзадачи по аргументу может быть сколь угодно большой. То есть для функций (в том числе и нелинейных), для которых значение L на данном квадрате достаточно мало, может оказаться допустимым вообще не решать вспомогательные задачи на разделяющих отрезках.

Замечание 4.5.6. Важно отметить, что схема доказательства основного результата позволяет учитывать не только погрешность решения вспомогательных одномерных задач, но и погрешность вычисления градиента в точках x_δ (см. пп. 3 и 4 доказательства теоремы). Точнее говоря, в точках x_δ вектор градиента функции может вычисляться с точностью Δ: если $\nabla f(x_\delta)$ — точная величина градиента, то нам доступен некоторый вектор $v(x_\delta)$, для которого

$$\|v(x_\delta) - \nabla f(x_\delta)\|_2 \leqslant \Delta.$$

В том случае, если вектор $v(x_\delta)$ направлен в ту же часть делимого квадрата, что и вектор $\nabla f(x_\delta)$, все рассуждения п. 3 доказательства теоремы 4.5.1 из [44] остаются неизменными, так как исключается та

же часть квадрата, что и при $\Delta = 0$. Если же это не так, то либо вектор $\nabla f(x_\delta)$ направлен в ту часть, где лежит x_* – точное решение задачи, а вектор $v(x_\delta)$ – в иную (очевидно, что в результате будет исключена часть, которой не принадлежит x_*, так что рассуждения в пункте 3 доказательства теоремы также остаются справедливыми), либо же вектор $v(x_\delta)$ в ту часть, где лежит x_*, а $\nabla f(x_\delta)$ — в иную.

В таком случае имеем треугольник, образованный векторами $\nabla f(x_\delta) - \nabla f(x_0)$ и $v(x_\delta) - \nabla f(x_\delta)$. Из неравенства треугольника следует, что если $\|\nabla f(x_\delta) - \nabla f(x_0)\|_2 \leqslant L\delta$ и $\|v(x_\delta) - \nabla f(x_\delta)\|_2 \leqslant \Delta$, то

$$\|v(x_\delta) - \nabla f(x_0)\|_2 \leqslant L\delta + \Delta.$$

Векторы $v(x_\delta)$ и $\nabla f(x_0)$ образуют тупоугольный треугольник, причём длина его наибольшей стороны $\|v(x_\delta) - \nabla f(x_0)\|_2 \leqslant L\delta + \Delta$. Следовательно, $\|v(x_\delta)\|_2 \leqslant L\delta + \Delta$. Из треугольника, образуемого векторами $v(x_\delta)$ и $\nabla f(x_\delta)$ ввиду неравенства треугольника, получаем

$$\|\nabla f(x_\delta)\|_2 \leqslant \|v(x_\delta)\|_2 + \|v(x_\delta) - \nabla f(x_\delta)\|_2 \leqslant L\delta + 2\Delta. \tag{4.56}$$

Далее, следуя схеме рассуждений доказательства теоремы 4.5.1, получим

$$(L\delta + 2\Delta) \cdot R\left(\sqrt{2} + \sqrt{5}\right)\left(1 - \frac{\varepsilon}{MR\sqrt{2}}\right) \leqslant \frac{\varepsilon}{2},$$

откуда

$$2\Delta + L\delta \leqslant \frac{\varepsilon}{2R\left(\sqrt{2} + \sqrt{5}\right)\left(1 - \dfrac{\varepsilon}{MR\sqrt{2}}\right)}. \tag{4.57}$$

Таким образом, метод можно применить при наличии двух типов погрешностей: при решении вспомогательных одномерных задач оптимизации и при вычислении (направлений) градиентов. Эти два типа погрешностей аккумулируют итоговую погрешность. При увеличении одной из погрешности можно уменьшать значение другой так, чтобы сохранялось последнее неравенство, гарантирующее достижение приемлемого качества решения.

Следующее замечание написано совместно с учащимся ФМЛ 239 Санкт-Петербурга Дмитрием Аркадьевичем Пасечнюком.

Замечание 4.5.7. Стоит отметить, что в случае, когда вектор-приближение градиента $v(x_\delta)$ направлен противоположно градиенту в

точке решения вспомогательной задачи $\nabla f(x_0)$, в силу полученной выше оценки (4.56)

$$\|\nabla f(x_\delta)\|_2 \leqslant L\delta + 2\Delta,$$

т.е. при достаточно близких к нулю малых величинах δ и Δ вектор $\nabla f(x_\delta)$ можно считать достаточно малым по модулю. С учетом условия (4.57) для суммарной погрешности решения вспомогательных задач приходим к следующей оценке значения градиента:

$$\|\nabla f(x_\delta)\|_2 \leqslant \frac{\varepsilon}{2R\left(\sqrt{2}+\sqrt{5}\right)\left(1-\dfrac{\varepsilon}{MR\sqrt{2}}\right)}.$$

В таком случае можно считать полученную точку x_δ достаточно хорошим приближением точного решения задачи. Действительно, используя выпуклость f и неравенство Коши-Буняковского, имеем

$$f(x_\delta) - f(x_*) \leqslant \|\nabla f(x_\delta)\|_2 \cdot \|x_* - x_\delta\|,$$

откуда можно вывести следующую оценку для неточности по функции при выборе такого решения:

$$f(x_\delta) - f(x_*) \leqslant \frac{\varepsilon R\sqrt{5}}{4R\left(\sqrt{2}+\sqrt{5}\right)\left(1-\dfrac{\varepsilon}{MR\sqrt{2}}\right)} \leqslant \frac{\varepsilon}{4\left(\sqrt{2}-\dfrac{\varepsilon}{MR}\right)}.$$

Откуда при дополнении условий теоремы 4.5.1 ограничением на M:

$$M \geqslant \frac{\varepsilon}{1.6R},$$

получаем

$$f(x_\delta) - f(x_*) \leqslant \varepsilon.$$

Отправляясь от следующего неравенства для неточного градиента

$$\|\nabla f(x_\delta)\|_2 \leqslant \|v(x_\delta)\|_2 + \Delta,$$

выпишем достаточное условие на вектор $v(x_\delta)$, при выполнении которого найденная точки будет приемлемой:

$$\|v(x_\delta)\|_2 \leqslant L\delta + \Delta.$$

Таким образом, при использовании данного наблюдения, критерий остановки метода можно переписать так:

$$\left(n = \log_2 \frac{2MR\sqrt{2}}{\varepsilon}\right) \vee \left(\|v(x_\delta)\| \leqslant L\delta + \Delta\right).$$

4.5.3 О применимости метода к задачам выпуклого программирования с двумя функционалами ограничений

Рассмотрим применимость предложенного подхода к условным задачам многомерной оптимизации с небольшим числом ограничений (которые за счёт операции максимизации можно свести к двум). Заметим, что выполнение условия Слейтера позволяет компактифицировать значения двойственных переменных и по сути рассматривать в качестве двойственной задачи двумерную задачу оптимизации на некотором квадрате. Кратко наметим схему рассуждений, применимых к рассматриваемым задачам выпуклого программирования вида (4.53).

Если для некоторого $x_\delta(\lambda)$ и $\lambda = (\lambda_1, \lambda_2)$ верно

$$|\lambda_1 g_1(x_\delta(\lambda)) + \lambda_2 g_2(x_\delta(\lambda))| \leqslant \varepsilon, \tag{4.58}$$

где $x_\delta(\lambda)$ — приближённое решение задачи

$$x(\lambda) = \arg\min_{x \in Q} \{f(x) + \lambda_1 g_1(x) + \lambda_2 g_2(x)\} \tag{4.59}$$

с точностью δ по функции, то

$$f(x_\delta(\lambda)) - f(x_*) \leqslant \varepsilon + \delta.$$

Действительно, пусть $\lambda_* = (\lambda_1^*, \lambda_2^*)$ — решение задачи (4.54) и $x_* = x(\lambda_*)$. Тогда

$$\begin{gathered} f(x_\delta(\lambda)) + \lambda_1 g_1(x_\delta(\lambda)) + \lambda_2 g_2(x_\delta(\lambda)) \leqslant \\ \leqslant f(x(\lambda)) + \lambda_1 g(x_\delta(\lambda)) + \lambda_2 g(x_\delta(\lambda)) + \delta = \varphi(\lambda_1, \lambda_2) + \delta \leqslant \\ \leqslant \varphi(\lambda_*) + \delta = f(x_*) + \lambda_1^* g(x_*) + \lambda_2^* g(x_*) + \delta \leqslant f(x_*) + \delta, \end{gathered}$$

откуда

$$f(x_\delta(\lambda)) - f(x_*) \leqslant -(\lambda_1 g_1(x_\delta(\lambda)) + \lambda_2 g_2(x_\delta(\lambda))) + \delta \leqslant \varepsilon + \delta,$$

что и требовалось.

Поэтому задачу (4.53) можно сводить к (4.54) и решать (4.54) методом Ю. Е. Нестерова на подходящем квадрате с критерием остановки (4.58). Если решать задачи (4.59) с точностью ε по функции и с точностью $\delta = O(\varepsilon)$ по аргументу, то можно показать для гладких f, g_1 и g_2 оценку скорости сходимости $O\left(\log^3 \frac{1}{\varepsilon}\right)$.

Заключительные замечания к главе 4

В первом разделе данной главы предложен метод градиентного типа с адаптацией к константе L гладкости и параметрам погрешностей на классе задач с условием градиентного доминирования. Обоснована линейная скорость сходимости метода для гладких задач с точностью до величины погрешностей. Отметим, что условие градиентного доминирования заведомо верно для сильно выпуклой целевой функции f. Однако довольно хорошо известны примеры, когда нельзя быть уверенным даже в выпуклости $f(x)$, но (PL)-условие имеет место [38] (раздел 4.3), [28], [160]. Например, рассмотрим систему нелинейных уравнений $g(x) = 0$, записанную в векторном виде, то есть $g : \mathbb{R}^n \to \mathbb{R}^m$, $m \leqslant n$ и рассмотрим задачу нахождения какого-нибудь решения этой системы. Если ввести матрицу Якоби отображения g : $\frac{\partial g(x)}{\partial x} = \left\| \frac{\partial g_i(x)}{\partial x_j} \right\|_{i,j=1}^{m,n}$ и предположить, что существует такое $\mu > 0$, что для всех $x \in \mathbb{R}^n$ имеет место равномерная невырожденность матрицы Якоби:

$$\lambda_{min} \left(\frac{\partial g(x)}{\partial x} \cdot \left[\frac{\partial g(x)}{\partial x} \right]^T \right) \geqslant \mu.$$

Тогда функция $f(x) = \|g(x)\|_2^2$ удовлетворяет условию (4.1) для произвольного x_* такого, что $f(x_*) = 0$, то есть $g(x_*) = 0$ [145]. Важно также отметить, использование подхода п. 4.1 даёт возможность оценивать качество найденного приближённого решения по значению неотрицательной целевой функции в начальной точке $f(x^0) \geqslant f(x^0) - f(x_*)$ без знания оценки расстояния от x^0 до точного решения, что выделяет данный подход по сравнению с известными в литературе (в т.ч. в других разделах настоящей работы) концепциями неточного оракула и может быть удобным при отсутствии априорного представления о том, где находится точное решение (например, на неограниченной допустимой области).

Далее, вводится аналог (δ, L, μ)-оракула в модельной общности. Предложен метод с адаптивной настройкой на параметр L, доказана оценка скорости сходимости. В следующем разделе главы описан вариант этого метода с адаптивной настройкой на параметры гладкости и погрешности. При этом, показана применимость введенных во втором и третьем разделах главы для концепции относительной гладкости и относительной сильной выпуклости целевого функционала. Отметим, что

удается обосновать сходимость не только функции, но и по аргументу. Отметим, что в разделе 4.3, где предложен метод с адаптивной настройкой на величину погрешности мы полагаем, что вспомогательные задачи решаются точно. На самом деле, если допустить погрешности $\widetilde{\delta}$ решения таких задач согласно (0.11), то можно доказать отсутствие их накопления в итоговых оценках скорости сходимости предложенных методов полностью аналогично рассуждениям, приведённым ранее в разделе 4.2.

В качестве актуальной задачи на будущее можно было бы выделить проблему более детального исследования так называемых ускоренных методов для рассмотренных в разделах 4.1, 4.2 и 4.3 классов задач. В частности, к ускоренным методам относят самые разные вариации так называемого быстрого градиентного метода (БГМ) (см. например [38, 90, 132]). Для задач выпуклой гладкой оптимизации без погрешностей БГМ гарантирует лучшую оценку скорости сходимости по сравнению с полученной. Известно, что в сильно выпуклом случае использование ускоренных методов позволяет уменьшить знаменатель геометрической прогрессии, которая описывает скорость сходимости. Более того, неадаптивные ускоренные методы для релаксаций условия сильной выпуклости исследовались в [132]. Однако стоит отметить, что при наличии погрешностей ситуация становится уже более нетривиальной: в отличие от обычного градиентного метода возможно их накопление в итоговой оценке [91], либо же необходимо использовать довольно ограничительные условия на величины таких погрешностей [89]. В этом плане можно отметить вариант ускоренного метода для минимизации функций, допускающих в произвольной запрошенной точке (δ, Δ, L, μ)-модель, рассмотренный в замечании 4.3.7 (алгоритм 14). В этом замечании как раз и показано, что преимущество использования именно ускоренного метода при $\Delta > 0$ уже не столь очевидно. Также заметим, что в силу недавно полученных нижних оценок [92], по-видимому, не возможно предложить ускоренный метод, который применим в общем случае для относительно гладких задач [124]. Представляется интересной задача исследования применимости результатов настоящей работы к бесконечномерным задачам, в частности к линейным и некоторым классам нелинейных операторных уравнений.

В заключении главы рассмотрены методы для задачи минимизации сильно выпуклого функционала выпуклой функции при одном или двух

выпуклых функциональных ограничениях. В этом случае, вообще говоря, задача не сводится к сильно выпукло-вогнутой седловой задаче. Суть предложенного подхода заключается в переходе к одномерной двойственной задаче и решении ее методом дихотомии для одного функционала ограничения или методом Ю. Е. Нестерова в случае двух функционалов ограничений. При этом в общем случае ввиду невозможности явного построения двойственной задачи необходимо учитывать погрешности при нахождении двойственной функции φ и ее производной. С учетом этого в работе получены оценки сложности предложенной алгоритмической схемы. Эти оценки близки к оптимальным в гладком случае, но проигрывают другим методам в негладком случае. Однако за счет адаптивности критерия остановки реально метод может приводить к достижению приемлемого качества решения быстрее, чем согласно найденным теоретическим оценкам. Это проиллюстрировано экспериментами (раздел 4.4.8), в ходе которых было проведено сравнение работы метода с так называемым универсальным методом для седловых задач.

Основные результаты настоящей главы опубликованы в работах [1, 44, 52–54, 168, 174].

ГЛАВА 5

Адаптивные методы зеркального спуска для задач оптимизации с выпуклыми функциональными ограничениями

Введение

Настоящая глава посвящена, в основном, адаптивным методам зеркального спуска для задач выпуклого программирования с выпуклыми липшицевыми функционалами ограничений. Важно, что при этом для некоторых из разработанных алгоритмических методов удаётся обосновать сохранение с точностью до умножения на константу оптимальных вычислительных гарантий (оценок скорости сходимости) для широкого класса квазивыпуклых гладких и негладких целевых функционалов, не удовлетворяющих даже условию Липшица.

Также рассматриваются методы, применимые к относительно липшицевым целевым функционалам и функционалам ограничений [125, 157] с сохрением оптимальных оценок скорости сходимости. Напомним, что относительная липшицевость введена недавно в [125] и по сути обобщает обычное условие Липшица. Точнее говоря, такое условие предполагает замену ограниченности нормы субградиента $\|\nabla f(x)\|_* \leqslant M_f$ следующим более мягким требованием (*относительной липшицевостью*):

$$\|\nabla f(x)\|_* \leqslant \frac{M_f\sqrt{2V(y,x)}}{\|y-x\|} \quad \forall x, y \in Q,\ y \neq x. \tag{5.1}$$

Важно, что при этом порождающая функция d уже не обязательно сильно выпукла. В работе [125] предложены детерминированный и стохастический методы зеркального спуска для задач минимизации выпуклого относительно липшицева целевого функционала. Мы несколько ослабляем условие относительной липшицевости [125], рассматривая вместо (5.1) следующее неравенство

$$\langle \nabla f(x), x-y\rangle \leqslant M_f\sqrt{2V(y,x)} \quad \forall x, y \in Q.$$

Для выделенного в настоящей главе класса задач даже в случае гладкости целевой функции задачи функционалы ограничений могут быть, вообще говоря, негладким (недифференцируемыми), в том числе и за счёт перехода к операции взятия максимума. Поэтому для выделенного класса задач вполне естественно рассматривать субградиентные методы. Исследования в области субградиентных методов для негладкой оптимизации восходят к известным пионерским работам, одна из которых посвящена градиентному методу для задач с евклидовым расстоянием [62], а другая — его обобщению для задач с функциональными ограничениями [48]. В работе [48] предложена идея переключения шагов между направлением субградиента целевого функционала и направлением субградиента функционала ограничения. Обобщение метода градиентного спуска для задач с неевклидовым расстоянием называют *методом зеркального спуска*. Этот метод был предложен в [36, 37] для задач без ограничений (см. также [73]). Зеркальный спуск для задач с функциональными ограничениями был предложен в [36] (см. также [71]).

В настоящей главе мы рассматриваем некоторые алгоритмы зеркального спуска для задач минимизации выпуклого функционала f на некотором выпуклом замкнутом множестве с ограничением, порожденным выпуклым, липшицевым (или относительно липшицевым) и не обязательно гладким функционалом $g(x) \leqslant 0$. Важно, что получены оценки скорости сходимости методов для целевых функционалов различного уровня гладкости. В частности, целевой функционал f может не удовлетворять свойству Липшица, но иметь липшицев градиент. Например, квадратичные функционалы не удовлетворяют обычному свойству Липшица (или константа Липшица достаточно большая), но имеют липшицев градиент. Можно рассматривать и негладкие выпуклые функции, равные максимуму конечного набора дифференцируемых функционалов с липшицевым градиентом. Например, если $A_i (i \in \overline{1,m})$ — положительно полуопределённые матрицы ($x^T A_i x \geqslant 0$ для всякого $x \in Q$; где Q — область определения задачи) и целевой функционал имеет вид

$$f(x) = \max_{i=\overline{1,m}} f_i(x), \tag{5.2}$$

где

$$f_i(x) = \frac{1}{2}\langle A_i x, x\rangle - \langle b_i, x\rangle + c_i, \quad i = \overline{1,m}. \tag{5.3}$$

для некоторых фиксированных $b_i \in \mathbb{R}^n$ и $c_i \in \mathbb{R}$, для всех $i = \overline{1,m}$. Отметим, что функционалы вида (5.2)–(5.3) естественно возникают в задачах проектирования механических конструкций Truss Topology Design со взвешенными балками [141]. Для задач минимизации функционалов такого типа при наличии выпуклых липшицевых ограничений в [58, 68, 169] на базе методики работ [139, 141] нами были предложены некоторые новые адаптивные алгоритмы зеркального спуска, а также обоснована их оптимальность. При этом функционалы ограничений в [139, 141] обязательно липшицевы. Однако методы с неадаптивными непродуктивными шагами применимы в случае относительно липшицевых функционалов ограничений (в частности, и для функционалов вида (5.2), (5.3)).

Перечислим основные результаты настоящей главы.

– Доказана оптимальность с точки зрения нижних оракульных оценок предложенных методов в [58, 68, 169] для задач с выпуклым липшицевым целевым функционалом при наличии выпуклых липшицевых функционалов ограничений. При этом обосновано сохранение оптимальных оценок скорости сходимости, ранее полученных на более узком классе задач выпуклого программирования с гёльдеровыми (суб)дифференцируемыми целевыми функционалами.

– С помощью рестартов (перезапусков) методов для выпуклых задач предложены новые алгоритмы зеркального спуска для задач минимизации μ-сильно выпуклого функционала f с неположительным, μ-сильно выпуклым и липшицевым негладким функционалом ограничения g. Заметим, что техника рестартов (перезапусков) метода для выпуклых задач с целью повышения теоретических вычислительных гарантий для сильно выпуклых задач восходит к работам 1980-х годов. Подходы такого типа были использованы, в частности, в [111] для обоснования более высокой скорости сходимости метода зеркального спуска для сильно выпуклого целевого функционала в задачах без ограничений.

– Показано, что некоторые из разработанных методов зеркального спуска применимы к задачам минимизации квазивыпуклого субдифференцируемого по Кларку целевого функционала с ненулевым субградиентом в любой точке, а также выпуклым липшицевым функциональным ограничением.

– Предложен новый адаптивный метод зеркального спуска для задач онлайн-оптимизации в случае выпуклых (возможно, негладких) лип-

шицевых целевых функционалов и нескольких выпуклых липшицевых функциональных ограничений. Обоснована оптимальность оценок скорости сходимости метода на рассмотренном классе задач.

– Разработаны частично адаптивные методы (с постоянными продуктивными и/или непродуктивными шагами, но с адаптивными критериями остановки) зеркального спуска для задач оптимизации с функциональными ограничениями в предположении относительной липшицевости целевого функционала и функционала ограничения. Как приложение, предложен метод для минимизации выпуклого однородного функционала с функциональными ограничениями, который гарантирует достижение заданной относительной точности приближённого решения при достаточно общих предположениях.

Приведем формальную постановку рассматриваемой задачи оптимизации, а также напомним необходимые вспомогательные понятия и результаты. Пусть $(E, \|\cdot\|)$ — конечномерное нормированное векторное пространство и E^* — сопряженное пространство к E со стандартной нормой:

$$\|y\|_* = \max_x \{\langle y, x\rangle, \|x\| \leqslant 1\},$$

где $\langle y, x\rangle$ — значение линейного непрерывного функционала y в точке $x \in E$.

Данная глава, в основном, посвящена следующему типу задач:

$$f(x) \to \min, \quad x \in Q, \tag{5.4}$$

$$g(x) \leqslant 0, \tag{5.5}$$

где f и g выпуклы и удовлетворяют соответствующим предположениям о гладкости. Если функционалов ограничений несколько $g_p(x) \leqslant 0$ ($p = \overline{1, M}$ для некоторого натурального M), то можно рассмотреть один функционал вида $g(x) = \max\limits_{p=\overline{1,M}} g_p(x)$. Сделаем предположение о разрешимости задачи (5.4)–(5.5). Обозначим через x_* — точное решение задачи (5.4)–(5.5). Наша цель — предложить метод, который позволяет найти за конечное число шагов ε-решение поставленной задачи $\widehat{x} \in Q$:

$$f(\widehat{x}) - f(x_*) \leqslant C_f \cdot \varepsilon \quad \text{при} \quad g(\widehat{x}) \leqslant C_g \cdot \varepsilon$$

для некоторых постоянных $C_f, C_g > 0$. Всюду далее мы предполагаем, что начальное приближение x^0 для всех методов выбирается так, что $d(x^0) = \min\limits_{x \in Q} d(x)$.

Всюду далее будем считать, что $Q \subset E$ — замкнутое выпуклое множество. Рассмотрим два выпуклых субдифференцируемых функционала f и $g : Q \to \mathbb{R}$. Если не оговорено иное, то предположим, что функционал g удовлетворяет условию Липшица относительно нормы $\|\cdot\|$, т. е. существует $M_g > 0$ такое, что

$$|g(x) - g(y)| \leqslant M_g \|x - y\|$$

для всяких $x, y \in Q$. Это означает, что во всякой точке $x \in Q$ существует субградиент $\nabla g(x)$, причём $\|\nabla g(x)\|_* \leqslant M_g$. Напомним, что для дифференцируемого функционала g субградиент $\nabla g(x)$ совпадает с обычным градиентом.

Отметим, что часть результатов работы (раздел 5.4) посвящена постановке задачи для μ-сильно выпуклых субдифференцируемых функционалов f и $g : Q \to \mathbb{R}$, т.е. для произвольных $x, y \in Q$ имеет место неравенство

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

и аналогичное неравенство верно для g с тем же параметром сильной выпуклости μ. Напомним стандартное определение оператора проектирования

$$\mathrm{Mirr}_x(p) = \arg\min_{u \in Q} \left\{ \langle p, u \rangle + V(u, x) \right\} \text{ для каждого } x \in Q \text{ и } p \in E^*.$$

Сделаем предположение о том, что оператор $\mathrm{Mirr}_x(p)$ легко вычислим.

Вспомним также известное утверждение (см., например [74]), на котором базируются многие дальнейшие рассуждения.

Лемма 5.0.1. *Пусть $f : Q \to \mathbb{R}$ — выпуклый субдифференцируемый функционал на выпуклом множестве Q и $z = Mirr_y(h\nabla f(y))$ для некоторого $y \in Q$ и $h > 0$. Тогда для произвольного $x \in Q$ справедливо неравенство*

$$h(f(y) - f(x)) \leqslant h\langle \nabla f(y), y - x \rangle \leqslant \frac{h^2}{2}||\nabla f(y)||_*^2 + V(x, y) - V(x, z).$$

Сформулируем и докажем базовую лемму для зеркальных спусков на более общем классе относительно липшицевых функционалов.

Лемма 5.0.2. *Пусть при* $h > 0$ $z = Mirr_y(h\nabla f(y))$, *а также верно неравенство*

$$\langle \nabla f(y), y - z\rangle \leqslant M_f\sqrt{2V(z,y)} \ \ \forall y, z \in Q. \tag{5.6}$$

Тогда для произвольных $x \in Q$ *и* $h > 0$ *верно неравенство*

$$h\langle \nabla f(y), y - x\rangle \leqslant \frac{h^2M_f^2}{2} + V(x,y) - V(x,z).$$

Доказательство. Поскольку $z = \mathrm{Mirr}_y(h\nabla f(y))$, то

$$\langle h\nabla f(y), z - x\rangle \leqslant \langle \nabla d(z) - \nabla d(y), x - z\rangle$$

для всякого $x \in Q$, и поэтому

$$\langle h\nabla f(y), y - x\rangle \leqslant \langle h\nabla f(y), y - z\rangle + \langle \nabla d(z) - \nabla d(y), x - z\rangle =$$

$$= \langle h\nabla f(y), y - z\rangle + (d(x) - d(y) - \langle \nabla d(y), x - y\rangle) -$$

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

$$= \langle h\nabla f(y), y - z\rangle + V(x,y) - V(x,z) - V(z,y) \leqslant$$

$$\leqslant \frac{h^2M_f^2}{2} + V(x,y) - V(x,z),$$

поскольку ввиду (5.6)

$$\langle h\nabla f(y), y - z\rangle \leqslant hM_f\sqrt{2V(z,y)} \leqslant \frac{h^2M_f^2}{2} + V(z,y),$$

т. е. $\langle h\nabla f(y), y - z\rangle - V(z,y) \leqslant \dfrac{h^2M_f^2}{2}$. □

Отметим, что платой за общность рассматриваемого класса функционалов будет замена нормы (суб)градиента f на константу Липшица M_f. По-видимому, это приводит к проблематичности построения методов с адаптивным шагом, связанным с делением на норму (суб)градиента.

5.1 Адаптивный и частично адаптивный алгоритмы зеркального спуска для задач с выпуклыми функционалами различного уровня гладкости. Случай гёльдерова целевого функционала

Перейдем теперь к описанию предложенных нами методов [68,169] для задач (5.4)–(5.5). Начнем со следующего адаптивного алгоритма зеркального спуска для задач (5.4)–(5.5) из ([68], п. 3.3). Выпишем этот метод для случая нескольких функционалов ограничений $g_p(x) \leqslant 0$ ($p = \overline{1, M}$ для некоторого натурального M).

Алгоритм 17 Адаптивный зеркальный спуск, деление на норму субградиента на продуктивном шаге.

Require: точность $\varepsilon > 0$; начальная точка $x^0; \Theta_0; Q; d(\cdot)$.

$I := \emptyset$

$N \leftarrow 0$

repeat

  if $g(x^N) \leqslant \varepsilon$ **then**

    $h_N \leftarrow \frac{\varepsilon}{\|\nabla f(x^N)\|_*}$

    $x^{N+1} \leftarrow Mirr_{x^N}(h_N \nabla f(x^N))$ ("продуктивные шаги")

    $N \to I$

  else

    $(g_{m(N)}(x^N) > \varepsilon) \to$ (для некоторого $m(N) \in \{1, \ldots, M\}$)

    $h_N \leftarrow \frac{\varepsilon}{\|\nabla g_{m(N)}(x^N)\|_*^2}$

    $x^{N+1} \leftarrow Mirr_{x^N}(h_N \nabla g_{m(N)}(x^N))$ ("непродуктивные шаги")

  end if

  $N \leftarrow N + 1$

until $\Theta_0^2 \leqslant \frac{\varepsilon^2}{2}\left(|I| + \sum\limits_{k \notin I} \frac{1}{\|\nabla g(x^k)\|_*^2}\right)$

Ensure: $\bar{x}^N := \arg\min\limits_{x^k, k \in I} f(x^k)$

Как видно из листинга приведённого алгоритма искомая точка выбирается среди точек x^k, для которых $g(x^k) \leqslant \varepsilon$. Поэтому будем назы-

вать *продуктивными* шаги, для которых $g(x^k) \leqslant \varepsilon$. Если же выполнено обратное неравенство $g(x^k) > \varepsilon$, то такой шаг k будем называть *непродуктивным*.

Аналогично [139] введем для целевого функционала f в точке $y \in Q$ следующую вспомогательную величину:

$$v_f(x,y) = \begin{cases} \left\langle \dfrac{\nabla f(x)}{\|\nabla f(x)\|_*}, x - y \right\rangle, & \nabla f(x) \neq 0 \\ 0, & \nabla f(x) = 0 \end{cases}, \quad x \in Q.$$

Мы допускаем, что в ходе работы метода можно использовать произвольный субградиент $\nabla f(x)$.

Для оценки скорости сходимости алгоритма 17 в [68] получен следующий результат.

Теорема 5.1.1. *Пусть известна константа $\Theta_0 > 0$ такая, что $d(x_*) \leqslant \Theta_0^2$. Если $\varepsilon > 0$ — фиксированное число, то алгоритм 17 работает не более*

$$N = \left\lceil \frac{2\max\{1, M_g^2\}\Theta_0^2}{\varepsilon^2} \right\rceil \tag{5.7}$$

итераций, причём после его остановки справедливо неравенство

$$\min_{k \in I} v_f(x^k, x_*) \leqslant \varepsilon, \ \max_{k \in I} g(x^k) \leqslant \varepsilon.$$

Доказательство. Пусть $[N] = \{k \in \overline{0, N-1}\}$, $J = [N] \setminus I$ — набор номеров непродуктивных шагов.

1) Для продуктивных шагов по лемме 5.0.1 имеем, что

$$h_k \langle \nabla f(x^k), x^k - x \rangle \leqslant \frac{h_k^2}{2} ||\nabla f(x^k)||_*^2 + V(x, x^k) - V(x, x^{k+1}).$$

Примем во внимание, что $\frac{h_k^2}{2} ||\nabla f(x^k)||_*^2 = \frac{\varepsilon^2}{2}$. Тогда

$$\begin{aligned} h_k \langle \nabla f(x^k), x^k - x_* \rangle = \varepsilon \left\langle \frac{\nabla f(x^k)}{||\nabla f(x^k)||_*}, x^k - x_* \right\rangle = \varepsilon v_f(x^k, x_*) \leqslant \\ \leqslant \frac{\varepsilon^2}{2} + V(x_*, x^k) - V(x_*, x^{k+1}). \end{aligned} \tag{5.8}$$

2) Аналогично для «непродуктивных» шагов $k \in J$ (под $g_m(k)$ мы понимаем любое ограничение, для которого $g_m(k) > \varepsilon$) по лемме 5.0.1:

$$h_k\left(g_{m(k)}(x^k)-g_{m(k)}(x_*)\right)\leqslant\frac{h_k^2}{2}||\nabla g_{m(k)}(x^k)||_*^2+V(x_*,x^k)-V(x_*,x^{k+1})=$$
$$=\frac{\varepsilon^2}{2||\nabla g_{m(k)}(x^k)||_*^2}+V(x_*,x^k)-V(x_*,x^{k+1}). \tag{5.9}$$

3) Из (5.8) и (5.9) при $x=x_*$ имеем:

$$\varepsilon\cdot\sum_{k\in I}v_f(x^k,x_*)+\sum_{k\in J}\frac{\varepsilon^2(g_{m(k)}(x^k)-g_{m(k)}(x_*))}{2||\nabla g_{m(k)}(x^k)||_*^2}\leqslant$$
$$\leqslant\frac{\varepsilon^2}{2}\cdot|I|+\sum_{k=0}^{N-1}(V(x_*,x^k)-V(x_*,x^{k+1})). \tag{5.10}$$

Заметим, что для любого $k\in J$

$$g_{m(k)}(x^k)-g_{m(k)}(x_*)\geqslant g_{m(k)}(x^k)>\varepsilon. \tag{5.11}$$

С учетом

$$\sum_{k=0}^{N-1}(V(x_*,x^k)-V(x_*,x^{k+1}))\leqslant\Theta_0^2$$

неравенство (5.10) может быть преобразовано следующим образом:

$$\varepsilon\sum_{k\in I}v_f(x^k,x_*)\leqslant|I|\cdot\frac{\varepsilon^2}{2}+\Theta_0^2-\sum_{k\in J}\frac{\varepsilon^2}{2||\nabla g_{m(k)}(x^k)||_*^2},$$

$$\sum_{k\in I}v_f(x^k,x_*)\geqslant|I|\min_{k\in I}v_f(x^k,x_*).$$

Таким образом,

$$\varepsilon\cdot\min_{k\in I}v_f(x^k,x_*)\cdot|I|\leqslant\frac{\varepsilon^2}{2}\cdot|I|+\Theta_0^2-\sum_{k\in J}\frac{\varepsilon^2}{2||\nabla g_{m(k)}(x^k)||_*^2}\leqslant\varepsilon^2\cdot|I|,$$

откуда

$$\min_{k\in I}v_f(x^k,x_*)\leqslant\varepsilon.$$

В завершении покажем, что $|I|\neq 0$. Предположим обратное: $|I|=0\Rightarrow|J|=N$, т.е. все шаги непродуктивны. Тогда с учётом (5.11) получаем, что

$$h_k(g_{m(k)}(x^k)-g_{m(k)}(x_*))>\frac{\varepsilon^2}{\|\nabla g_{m(k)}(x^k)\|_*^2}$$

и

$$\sum_{k=0}^{N-1} h_k(g_{m(k)}(x^k) - g_{m(k)}(x_*)) \leqslant \sum_{k=0}^{N-1} \frac{\varepsilon^2}{2\|\nabla g_{m(k)}(x^k)\|_*^2} + \Theta_0^2 \leqslant$$

$$\leqslant \sum_{k=0}^{N-1} \frac{\varepsilon^2}{\|\nabla g_{m(k)}(x^k)\|_*^2}.$$

Итак, получили противоречие. Это означает, что $|I| \neq 0$.

Отметим, что в силу условия Липшица для функционального ограничения на любой итерации работы алгоритма 17 справедливо неравенство $||\nabla g_{m(k)}(x^k)||_* \leqslant M_g$. Поэтому критерий остановки алгоритма 17 будет заведомо выполнен не более, чем после (5.7) итераций работы. □

Можно предложить также и частично адаптивный метод для задачи (5.4)–(5.5) [169]. Его отличие от алгоритма 17 в том, что адаптивно выбирается шаг лишь на продуктивных итерациях и критерий остановки неадаптивен. Мы выписываем этот метод уже в случае только одного функционала ограничения $g(x)$.

Аналогично теореме 5.1.1 проверяется следующий результат (см. также [169]).

Теорема 5.1.2. *Пусть $\varepsilon > 0$ — фиксированное число и алгоритм 18 работает*

$$N = \left\lceil \frac{2M_g^2\Theta_0^2}{\varepsilon^2} \right\rceil$$

итераций. Тогда

$$\min_{k \in I} v_f(x^k, x_*) \leqslant \frac{\varepsilon}{M_g}, \ \max_{k \in I} g(x^k) \leqslant \varepsilon.$$

Теперь рассмотрим конкретные оценки скорости сходимости рассмотренных методов, которые обосновывают их оптимальность с точки зрения теории оракульных оценок, восходящей к известной монографии А. С. Немировского и Д. Б. Юдина [74]. Точнее говоря, ввиду липшицевости и, вообще говоря, негладкости функционалов ограничений для оптимальности метода с точки зрения нижних оракульных оценок этого достаточно показать [74], что для достижения требуемой точности ε решения задачи (5.4)– (5.5) для каждого из рассмотренных в данном разделе статьи класса целевых функционалов достаточно $O\left(\varepsilon^{-2}\right)$ итераций метода, предполагающих вычисление (суб)градиента целевого функционала или ограничения.

Алгоритм 18 Частично адаптивная версия алгоритма 17

Require: точность $\varepsilon > 0$; начальная точка $x^0; \Theta_0; Q; d(\cdot)$.

$x^0 = \operatorname{argmin}_{x \in Q} d(x)$
$I := \emptyset$
$N \leftarrow 0$
repeat
  if $g(x^N) \leqslant \varepsilon \rightarrow$ **then**
    $h_N \leftarrow \frac{\varepsilon}{M_g \cdot \|\nabla f(x^N)\|_*}$
    $x^{N+1} \leftarrow Mirr_{x^N}(h_N \nabla f(x^N))$ ("продуктивные шаги")
    $N \rightarrow I$
  else
    $(g(x^N) > \varepsilon)$
    $h_N \leftarrow \frac{\varepsilon}{M_g^2}$
    $x^{N+1} \leftarrow Mirr_{x^N}(h_N \nabla g(x^N))$ ("непродуктивные шаги")
  end if
  $N \leftarrow N + 1$
until $N \geqslant \left\lceil \frac{2M_g^2 \Theta_0^2}{\varepsilon^2} \right\rceil$

Ensure: $\bar{x}^N := \underset{x^k, k \in I}{\arg\min} f(x^k)$

Лемма 5.1.3. *Введём следующую функцию:*

$$\omega(\tau) = \max_{x \in Q} \{f(x) - f(x_*) : ||x - x_*|| \leqslant \tau\},$$

где τ положительное число. Тогда для всякого $y \in Q$

$$f(y) - f(x_*) \leqslant \omega(v_f(y, x_*)).$$

На базе леммы 5.1.3 в работе [169] показано, как с использованием предыдущего утверждения и теоремы 5.1.2, можно оценить скорость сходимости алгоритма 18, если целевой функционал f дифференцируем и его градиент удовлетворяет условию Липшица:

$$||\nabla f(x) - \nabla f(y)||_* \leqslant L||x - y|| \quad \forall x, y \in Q.$$

Тогда для произвольного $x \in Q$ верно следующее неравенство [139]

$$f(x) \leqslant f(x_*) + ||\nabla f(x_*)||_* ||x - x_*|| + \frac{1}{2} L ||x - x_*||^2,$$

откуда

$$\min_{k \in I} f(x^k) - f(x_*) \leqslant \min_{k \in I} \left\{ ||\nabla f(x_*)||_* ||x^k - x_*|| + \frac{1}{2} L ||x^k - x_*||^2 \right\},$$

$$\min_{k \in I} f(x^k) - f(x_*) \leqslant ||\nabla f(x_*)||_* \frac{\varepsilon}{M_g} + \frac{L}{2} \frac{\varepsilon^2}{M_g^2}.$$

Последнее неравенство позволяет сформулировать следующий результат [169], причём для некоторого класса, вообще говоря, негладких целевых функционалов.

Следствие 5.1.4. *Предположим, что $f(x) = \max\limits_{i = \overline{1,m}} f_i(x)$, где f_i дифференцируемы в каждой точке $x \in Q$ и*

$$||\nabla f_i(x) - \nabla f_i(y)||_* \leqslant L_i ||x - y|| \quad \forall x, y \in Q.$$

Тогда

- *после*

$$N = \left\lceil \frac{2 \max\{1, M_g^2\} \Theta_0^2}{\varepsilon^2} \right\rceil$$

шагов работы алгоритма 17 будет верно следующее неравенство:

$$\min_{k \in I} f(x^k) - f(x_*) \leqslant \|\nabla f(x_*)\| \varepsilon + \frac{L}{2} \varepsilon^2;$$

- *после*

$$N = \left\lceil \frac{2M_g^2\Theta_0^2}{\varepsilon^2} \right\rceil$$

шагов работы алгоритма 18 будет справедлива оценка:

$$\min_{k\in I} f(x^k) - f(x_*) \leqslant ||\nabla f(x_*)||_* \frac{\varepsilon}{M_g} + \frac{L}{2}\frac{\varepsilon^2}{M_g^2},$$

где $L = \max\limits_{i=\overline{1,m}} L_i$.

Замечание 5.1.5. Мы рассматриваем условные задачи и поэтому совсем не обязательно, что $||\nabla f(x_*)||_* = 0$.

Опишем теперь оценки скорости сходимости алгоритмов 17 и 18 для классов целевых функционалов, не рассмотренных в [169].

Замечание 5.1.6. Пусть целевой функционал $f : Q \to \mathbb{R}$ удовлетворяет условию Липшица:

$$|f(x) - f(y)| \leqslant M_f||x - y|| \quad \forall x, y \in Q. \tag{5.12}$$

Тогда для произвольного $x \in Q$ верно неравенство

$$f(x) \leqslant f(x_*) + M_f||x - x_*||,$$

откуда

$$\min_{k\in I} f(x^k) - f(x_*) \leqslant \min_{k\in I} M_f \left\{||x^k - x_*||\right\}.$$

Поэтому имеет место следующий результат.

Следствие 5.1.7. *Пусть f удовлетворяет условию Липшица* (5.20) *на Q. Тогда*

- *после*

$$N = \left\lceil \frac{2\max\{1, M_g^2\}\Theta_0^2}{\varepsilon^2} \right\rceil$$

шагов работы алгоритма 17 будет верно следующее неравенство:

$$\min_{k\in I} f(x^k) - f(x_*) \leqslant M_f\varepsilon;$$

- *после*

$$N = \left\lceil \frac{2M_g^2\Theta_0^2}{\varepsilon^2} \right\rceil$$

шагов работы алгоритма 18 будет справедлива оценка:

$$\min_{k\in I} f(x^k) - f(x_*) \leqslant \frac{M_f}{M_g}\varepsilon.$$

Замечание 5.1.8. Если выбирать непродуктивные шаги постоянными $h_k^g = \frac{\varepsilon}{M_g^2}$, то нетрудно проверить, что полученные оценки скорости сходимости для соответствующей модификации алгоритма 17 сохранятся и в случае M_g-относительно липшицева функционала ограничения:

$$\langle g(y), y - x\rangle \leqslant M_g\sqrt{2V(y,x)}\ \forall x, y \in Q.$$

Замечание 5.1.9. Покажем также, как получить оптимальные оценки скорости сходимости рассматриваемых методов на классе гёльдеровых выпуклых целевых функционалов. Итак, пусть f удовлетворяет условию Гёльдера для некоторого $\nu \in [0;1)$:

$$|f(x) - f(y)| \leqslant M_\nu\|x - y\|^\nu \quad \forall x, y \in Q. \tag{5.13}$$

Например, можно рассматривать $f(x) = \sqrt{x}$ или $f(x) = \sqrt[4]{x}$.

Напомним следующее неравенство (см. лемму 2.2.3)

$$M_\nu a^\nu \leqslant M_\nu\left(\frac{M_\nu}{\delta}\right)^{\frac{1-\nu}{1+\nu}}\frac{a^2}{2} + \delta,$$

которое верно для каждого $\delta > 0$. Тогда, учитывая (5.13), имеем

$$|f(x) - f(y)| \leqslant \frac{M_\nu^{\frac{2}{1+\nu}}}{2\delta^{\frac{1-\nu}{1+\nu}}}\|x - y\|^2 + \delta.$$

Пусть $\delta = \varepsilon$. Тогда

$$|f(x) - f(y)| \leqslant \frac{M_\nu^{\frac{2}{1+\nu}}}{2\varepsilon^{\frac{1-\nu}{1+\nu}}}\|x - y\|^2 + \varepsilon.$$

Тогда после остановки алгоритма 17 неравенство $\min\limits_{k\in I} v_f(x^k, x_*) < \varepsilon$ приводит к следующей оценке качества найденного приближённого решения

$$f(\widehat{x}) - f^* \leqslant \frac{M_\nu^{\frac{2}{1+\nu}}}{2\varepsilon^{\frac{1-\nu}{1+\nu}}}\varepsilon^2 + \varepsilon = \frac{M_\nu^{\frac{2}{1+\nu}}}{2}\varepsilon^{1+\frac{2\nu}{1+\nu}} + \varepsilon. \tag{5.14}$$

Обратим внимание, что для $\varepsilon < 1$ неравенство (5.14) означает, что

$$f(\widehat{x}) - f^* \leqslant \widehat{M}\varepsilon$$

для некоторой постоянной

$$\widehat{M} = \frac{M_\nu^{\frac{2}{1+\nu}}}{2} + 1 > 0. \tag{5.15}$$

Таким образом, для задач с выпуклым гёльдеровым (суб)дифференцируемым целевым функционалом и выпуклыми липшицевым функциональным ограничением возможно достичь ε-решения после не более, чем $O\left(\varepsilon^{-2}\right)$ итераций алгоритма 17. Очевидно, что эта оценка является оптимальной с точностью до умножения на константу, поскольку она оптимальна даже на более узком классе липшицевых функционалов.

5.2 Алгоритмы зеркального спуска для условия проверки продуктивности, связанного с нормой субградиента ограничения в текущей точке

В этом пункте мы рассмотрим еще два варианта адаптивных зеркальных спусков с переключениями [51, 176]. Для данных методов несколько по-иному проверяется продуктивность шага. Как будет показывают результаты расчетов (см. раздел 5.6), эти подходы могут приводить к более эффективной работе в случае больших норм субградиентов функционалов ограничений. Пусть задана фиксированная точность $\varepsilon > 0$, начального приближения x^0 и некоторой величины Θ_0 такой, что $V(x_*, x^0) \leqslant \Theta_0^2$. Для алгоритма 19 справедлива следующая теорема.

Теорема 5.2.1. *После остановки предложенного алгоритма 19 для*

$$\widehat{x} = \frac{1}{\sum_{k\in I} h_k} \sum_{k\in I} h_k x^k$$

верно $f(\widehat{x}) - f(x_*) \leqslant \varepsilon$ *и* $g(\widehat{x}) \leqslant \varepsilon M_g$.

Алгоритм 19 Адаптивный зеркальный спуск, деление на квадрат нормы субградиента для непродуктивного шага.

Require: $\varepsilon > 0, \Theta_0 : d(x_*) \leqslant \Theta_0^2$

$x^0 = argmin_{x \in Q} d(x)$

$I =: \emptyset$

$N \leftarrow 0$

repeat

  if $g(x^N) \leqslant \varepsilon \|\nabla g(x^N)\|_*$ **then**

    $M_N = \|\nabla f(x^N)\|_*$, $h_N = \frac{\varepsilon}{M_N^2}$

    $x^{N+1} = Mirr_{x^N}(h_N \nabla f(x^N))$ // *"продуктивные шаги"*

    $N \to I$

  else

    $M_N = \|\nabla g(x^N)\|_*$, $h_N = \frac{\varepsilon}{M_N}$

    $x^{N+1} = Mirr_{x^N}(h_N \nabla g(x^N))$ // *"непродуктивные шаги"*

  end if

  $N \leftarrow N + 1$

until $2\frac{\Theta_0^2}{\varepsilon^2} \leqslant \sum_{j \in I} \frac{1}{M_j^2} + N - |I|$

Ensure: $\widehat{x} = \frac{1}{\sum_{k \in I} h_k} h_k x^k$

Доказательство. 1) Если $k \in I$, то

$$\begin{aligned} & h_k \left(f(x^k) - f(x_*) \right) \leqslant h_k \left\langle \nabla f(x^k), x^k - x_* \right\rangle \leqslant \\ & \leqslant \frac{h_k^2}{2} ||\nabla f(x^k)||_*^2 + V(x_*, x^k) - V(x_*, x^{k+1}) = \\ & = \frac{\varepsilon^2}{2} \cdot \frac{1}{||\nabla f(x^k)||_*^2} + V(x_*, x^k) - V(x_*, x^{k+1}). \end{aligned} \tag{5.16}$$

2) Если $k \notin I$, то $\frac{g(x^k)}{||\nabla g(x^k)||_*} > \varepsilon$ и $\frac{g(x^k) - g(x_*)}{||\nabla g(x^k)||_*} \geqslant \frac{g(x^k)}{||\nabla g(x^k)||_*} > \varepsilon$. Поэтому верны неравенства

$$\begin{aligned} & \varepsilon^2 < h_k \left(g(x^k) - g(x_*) \right) \leqslant \frac{h_k^2}{2} ||\nabla g(x^k)||_*^2 + \\ & + V(x_*, x^k) - V(x_*, x^{k+1}) = \frac{\varepsilon^2}{2} + V(x_*, x^k) - V(x_*, x^{k+1}), \text{ или} \\ & \frac{\varepsilon^2}{2} < V(x_*, x^k) - V(x_*, x^{k+1}). \end{aligned} \tag{5.17}$$

3) После суммирования неравенств (5.16) и (5.17) имеем:

$$\sum_{k \in I} h_k \left(f(x^k) - f(x_*) \right) \leqslant$$

$$\leqslant \sum_{k \in I} \frac{\varepsilon^2}{2||\nabla f(x^k)||_*^2} - \frac{\varepsilon^2 |J|}{2} + V(x_*, x^0) - V(x_*, x^{k+1}) =$$

$$= \frac{\varepsilon}{2} \sum_{k \in I} h_k - \frac{\varepsilon^2 |J|}{2} + \Theta_0^2 =$$

$$= \varepsilon \sum_{k \in I} h_k - \frac{\varepsilon^2}{2} \left(\sum_{k \in I} \frac{1}{||\nabla f(x^k)||_*^2} + |J| \right) + \Theta_0^2.$$

После выполнения критерия остановки алгоритма 19 будет верно

$$\sum_{k \in I} h_k \left(f(x^k) - f(x_*) \right) \leqslant \varepsilon \sum_{k \in I} h_k,$$

откуда для $\widehat{x} := \sum\limits_{k \in I} \frac{h_k x^k}{\sum_{k \in I} h_k}$

$$f(\widehat{x}) - f(x_*) \leqslant \varepsilon.$$

При этом $\forall k \in I\ g(x^k) \leqslant \varepsilon ||\nabla g(x^k)||_* \leqslant \varepsilon M_g$, откуда

$$g(\widehat{x}) \leqslant \frac{1}{\sum_{k\in I} h_k} \sum_{k\in I} h_k g(x^k) \leqslant \varepsilon M_g.$$

Остается лишь показать, что множество продуктивных шагов I непусто. Если $I = \emptyset$, то $|J| = N$ и п. 14 листинга означают, что $N \geqslant \frac{2\Theta_0^2}{\varepsilon^2}$. С другой стороны, из (5.17):

$$\frac{\varepsilon^2 N}{2} < V(x_*, x^0) \leqslant \Theta_0^2,$$

то есть получили противоречие и $I \neq \emptyset$. Теорема доказана □

Замечание 5.2.2. Оценим количество итераций, необходимых для выполнения критерия остановки алгоритма 19 в случае M_f-липшицевости целевого функционала. Ясно, что $\forall k \in I\ \|\nabla f(x^k)\|_* \leqslant M_f$ и поэтому

$$|J| + \sum_{k\in I} \frac{1}{\|\nabla f(x^k)\|_*^2} \geqslant |J| + \frac{|I|}{M_f^2} \geqslant (|I| + |J|)\frac{1}{\max\{1, M_f^2\}}.$$

Это означает, что при

$$N \geqslant \frac{2\Theta_0^2 \max\{1, M_f^2\}}{\varepsilon^2} \tag{5.18}$$

критерий остановки заведомо выполнен, то есть искомая точность достигается за $O\left(\varepsilon^{-2}\right)$ итераций, что указывает на оптимальность предложенного метода с точки зрения теории нижних оракульных оценок.

Замечание 5.2.3. Ввиду адаптивности алгоритм 19 применим и для целевых функционалов, не удовлетворяющих условию Липшица. Однако в этом случае уже нельзя гарантировать оптимальные оценки сложности.

Предложена ещё одна алгоритмическая схема с аналогичной проверкой продуктивности шага, но с фиксированным числом шагов. Её преимущество — возможность получения оптимальных оценок для функционалов различного уровня гладкости по аналогии с алгоритмами 17 и 18.

Заметим, что алгоритм 20 всегда работает фиксированное число шагов

$$N = \left\lceil \frac{2\Theta_0^2}{\varepsilon^2} \right\rceil \tag{5.19}$$

Аналогично теореме 5.1.1 проверяется следующий результат.

Алгоритм 20 Адаптивный зеркальный спуск, фиксированное число итераций.

Require: $\varepsilon > 0, \Theta_0 : d(x_*) \leqslant \Theta_0^2$

$x^0 = argmin_{x \in Q} d(x)$

$I =: \emptyset$

$N \leftarrow 0$

repeat

  if $g(x^N) \leqslant \varepsilon \|\nabla g(x^N)\|_*$ **then**

    $M_N = \|\nabla f(x^N)\|_*$, $h_N = \frac{\varepsilon}{M_N}$

    $x^{N+1} = Mirr_{x^N}(h_N \nabla f(x^N))$ // *"продуктивные шаги"*

    $N \to I$

  else

    $M_N = \|\nabla g(x^N)\|_*$, $h_N = \frac{\varepsilon}{M_N}$

    $x^{N+1} = Mirr_{x^N}(h_N \nabla g(x^N))$ // *"непродуктивные шаги"*

  end if

  $N \leftarrow N + 1$

until $2\frac{\Theta_0^2}{\varepsilon^2} \leqslant N$

Ensure: $\bar{x}^N := argmin_{x^k,\, k \in I} f(x^k)$

Теорема 5.2.4. *Пусть $\varepsilon > 0$ — фиксированное число и выполнен критерий остановки алгоритма 20. Тогда*

$$\min_{k \in I} v_f(x^k, x_*) \leqslant \varepsilon, \quad \max_{k \in I} g(x^k) \leqslant M_g \varepsilon.$$

Покажем, как на базе предыдущей теоремы получить оценку качества решения по функции. Важно, что при этом возможен учёт различного уровня гладкости целевого функционала.

Следствие 5.2.5. *Пусть f удовлетворяет условию Липшица*

$$|f(x) - f(y)| \leqslant M_f \|x - y\| \quad \forall x, y \in Q \tag{5.20}$$

на Q. Тогда после остановки алгоритма 20 будет верно следующее неравенство:

$$\min_{k \in I} f(x^k) - f(x_*) \leqslant M_f \varepsilon;$$

Подобно замечанию 5.1.9 можно показать оптимальность оценок скорости сходимости алгоритма 20 на классе гёльдеровых целевых

функционалов [176]. Также по аналогии с алгоритмом 17 на «непродуктивных» шагах ($k \notin I$) можно вместо субградиента ограничения maxтипа $g(x) = \max\limits_{m=\overline{1,M}} g_m(x)$ использовать субградиент любого из функционалов g_m, для которого верно $g_m(x^k) > \varepsilon \cdot \|\nabla g_m(x^k)\|$ (см. алгоритмы 21 и 22).

Алгоритм 21 Модификация алгоритма 19, много ограничений.

Require: $\varepsilon > 0, \Theta_0 : d(x_*) \leqslant \Theta_0^2$

$x^0 = argmin_{x \in Q} d(x)$

$I =: \emptyset$

$N \leftarrow 0$

repeat

  if $g(x^N) \leqslant \varepsilon \|\nabla g_{m(N)}(x^N)\|_*$ **then**

    $M_N = \|\nabla f(x^N)\|_*$, $h_N = \frac{\varepsilon}{M_N^2}$

    $x^{N+1} = Mirr_{x^N}(h_N \nabla f(x^N))$ // *"продуктивные шаги"*

    $N \to I$

  else

    // $(g_{m(N)}(x^N) > \varepsilon \|\nabla g_{m(N)}(x^N)\|_*)$ *для некоторого* $m(N) \in \{1, \ldots, M\}$

    $M_N = \|\nabla g_{m(N)}(x^N)\|_*$, $h_N = \frac{\varepsilon}{M_N}$

    $x^{N+1} = Mirr_{x^N}(h_N \nabla g_{m(N)}(x^N))$ // *"непродуктивные шаги"*

  end if

  $N \leftarrow N + 1$

until $2\frac{\Theta_0^2}{\varepsilon^2} \leqslant \sum_{j \in I} \frac{1}{M_j^2} + N - |I|$

Ensure: $\widehat{x} = \frac{1}{\sum_{k \in I} h_k} h_k x^k$

5.3 Оптимальность зеркальных спусков для условных задач с квазивыпуклыми целевыми функционалами

В данном разделе мы покажем, как можно обобщить полученные оценки скорости сходимости для части рассмотренных методов на некоторый класс квазивыпуклых локально липшицевых функционалов $f\colon Q \to \mathbb{R}$ ($Q \subset \mathbb{R}^n$). Напомним, что функционал $f\colon Q \to \mathbb{R}$ называют

Алгоритм 22 Модификация алгоритма 20, много ограничений.

Require: $\varepsilon > 0, \Theta_0 : d(x_*) \leqslant \Theta_0^2$

$x^0 = argmin_{x \in Q} d(x)$

$I =: \emptyset$

$N \leftarrow 0$

repeat

  if $g(x^N) \leqslant \varepsilon \|\nabla g_{m(N)}(x^N)\|_*$ **then**

    $M_N = \|\nabla f(x^N)\|_*$, $h_N = \frac{\varepsilon}{M_N}$

    $x^{N+1} = Mirr_{x^N}(h_N \nabla f(x^N))$ // *"продуктивные шаги"*

    $N \to I$

  else

    // $(g_{m(N)}(x^N) > \varepsilon \|\nabla g_{m(N)}(x^N)\|_*)$ *для некоторого* $m(N) \in \{1, \ldots, M\}$

    $M_N = \|\nabla g_{m(N)}(x^N)\|_*$, $h_N = \frac{\varepsilon}{M_N}$

    $x^{N+1} = Mirr_{x^N}(h_N \nabla g_{m(N)}(x^N))$ // *"непродуктивные шаги"*

  end if

  $N \leftarrow N + 1$

until $2\frac{\Theta_0^2}{\varepsilon^2} \leqslant N$

Ensure: $\widehat{x} = \frac{1}{\sum_{k \in I} h_k} h_k x^k$

квазивыпуклым, если:

$$f((1-t)x+ty) \leqslant \max\{f(x), f(y)\} \quad \forall t \in [0;1] \quad \forall x, y \in Q.$$

Введем класс негладких квазивыпуклых функционалов, допускающих аналоги оценок полученных оценок скорости сходимости. Будем считать, что для функционала f в любой точке области определения существует компактный субдифференциал Кларка $\partial_{Cl} f(x)$. Напомним это понятие ([30], § 2.2). Пусть $x_0 \in \mathbb{R}^n$ — фиксированная точка и $h \in \mathbb{R}^n$ — фиксированное направление. Положим

$$f^{\uparrow}_{Cl}(x_0; h) = \limsup_{x' \to x_0,\ \alpha \downarrow 0} \frac{1}{\alpha} \left[f(x' + \alpha h) - f(x') \right] \ .$$

Величина $f^{\uparrow}_{Cl}(x_0; h)$ называется верхней производной Кларка функционала f в точке x_0 по направлению h. Как известно, функция $f^{\uparrow}_{Cl}(x_0; h)$ субаддитивна и положительно однородна по h ([30], с. 17 – 18). Это обстоятельство позволяет определить субдифференциал функционала f в точке x_0 как следующее множество:

$$\partial_{Cl} f(x_0) := \{ v \in \mathbb{R} \quad | \quad f^{\uparrow}_{Cl}(x_0; g) \geqslant vg \quad \forall g \in \mathbb{R} \} \ , \tag{5.21}$$

то есть как субдифференциал выпуклого по h функционала $f^{\uparrow}_{Cl}(x_0; h)$ в точке $h = 0$ в смысле выпуклого анализа. Таким образом, по определению

$$f^{\uparrow}_{Cl}(x_0; h) = \max_{v \in \partial_{Cl} f(x_0)} \langle v, h \rangle \ . \tag{5.22}$$

Будем говорить, что функционал f субдифференцируем по Кларку в точке x_0, если множество $\partial_{Cl} f(x_0)$ непусто и компактно. В частности, если функция f локально липшицева, то она является субдифференцируемой по Кларку в любой точке области определения. Отметим, что для выпуклых функций субдифференциал Кларка совпадает с обычным субдифференциалом в смысле выпуклого анализа [30].

Напомним одно известное утверждение, которое вытекает из обычного неравенства Коши-Буняковского, а также $2ab \leqslant a^2 + b^2$. Поскольку функциональные ограничения у нас по-прежнему выпуклы, мы рассмотрим также отдельно оценку в выпуклом случае [74].

Лемма 5.3.1. *Пусть $f : Q \to \mathbb{R}$ — некоторый функционал. Для произвольного $y \in Q$, вектора $p_y \in E^*$ и некоторого $h > 0$ положим*

$z = Mirr_y(h \cdot p_y)$. *Тогда для произвольного* $x \in Q$

$$h\langle p_y, y - x\rangle \leqslant \frac{h^2}{2}||p_y||_*^2 + V(y,x) - V(z,x).$$

Для выпуклого субдифференцируемого (уже в обычном смысле) в точке y *функционала* f *предыдущее неравенство для произвольного субградиента* $p_y = \nabla f(y)$ *примет вид*

$$h \cdot (f(y) - f(x)) \leqslant \langle \nabla f(y), y - x\rangle \leqslant \frac{h^2}{2}||\nabla f(y)||_*^2 + V(y,x) - V(z,x).$$

Будем рассматривать алгоритм адаптивного зеркального спуска для рассматриваемых задач. Отметим, что ввиду предположения локальной липшицевости квазивыпуклого целевого функционала все его субградиенты конечны. Сделаем дополнительное предположение об отсутствии точек перегиба, т.е. градиент f может быть нулевым только в точке x_*.

Для оценки скорости сходимости этого метода подобно ([139], п. 3.2.2), для всякого ненулевого конечного субградиента (элемента субдифференциала Кларка) $\nabla f(x)$ целевого квазивыпуклого функционала f введём следующую вспомогательную величину

$$v_f(x,y) = \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|_*}, x - y \right\rangle, \quad x \in Q.$$

Если $\nabla f(x) = 0$, то положим $v_f(x,y) := 0$. Сделаем также предположение о том, что $\nabla f(x) \neq 0$ при $x \neq x_*$.

Аналогично теореме 5.1.1 с использованием леммы 5.3.1 проверяется следующая

Теорема 5.3.2. *Пусть* $\varepsilon > 0$ *— фиксированное число и выполнен критерий остановки алгоритма 17. Тогда*

$$\min_{k \in I} v_f(x^k, x_*) \leqslant \varepsilon.$$

Отметим, что алгоритм 17 работает не более

$$N = \left\lceil \frac{2\max\{1, M_g^2\}\Theta_0^2}{\varepsilon^2} \right\rceil$$

итераций.

Теперь покажем, как можно оценивать скорость сходимости предлагаемого метода. Для этого полезно следующее вспомогательное утверждение, которое есть аналог ([139], лемма 3.2.1). Напомним, что под x_* мы понимаем точное решение задачи (5.4)–(5.5).

Теорема 5.3.3. *Пусть $f : Q \to \mathbb{R}^n$ — локально липшицев квазивыпуклый функционал. Введем следующую функцию:*

$$\omega(\tau) = \max_{x \in Q} \{f(x) - f(x_*) : ||x - x_*|| \leqslant \tau\},$$

где τ - положительное число. Тогда для всякого $x \in Q$

$$f(x) - f(x_*) \leqslant \omega(v_f(x, x_*)).$$

Доказательство. Мы отправляемся от схемы рассуждений ([139], лемма 3.2.1) с тем отличием, что вместо обычного субдифференциала выпуклой функции будет использоваться субдифференциал Кларка. Можно проверить, что

$$v_f(x, x_*) = \min_y \{||y - x_*|| : \langle \nabla f(x), y - x \rangle = 0\}.$$

Действительно, пусть минимум выражения $||y - x_*||$ достигается при $y = y_*$: $\langle \nabla f(x), y_* - x \rangle = 0$. Тогда $\nabla f(x) = \lambda s$, где $\langle s, y_* - x_* \rangle = ||y_* - x_*||$ для некоторого s, причем $||s||_* = 1$. Поэтому

$$0 = \langle \nabla f(x), y_* - x \rangle = \lambda \langle s, y_* - x_* \rangle + \langle \nabla f(x), x_* - x \rangle,$$

откуда

$$\lambda = \frac{\langle \nabla f(x), x - x_* \rangle}{||y_* - x_*||} = ||\nabla f(x)||_*,$$

т.е. $v_f(x, x_*) = ||y_* - x_*||$.

Далее, из (5.21), (5.22) имеем: $f^{\uparrow}_{Cl}(x, h) = \max\limits_{\nabla f(x) \in \partial_{Cl} f(x_0)} \langle \nabla f(x), h \rangle$. Для всякого направления h, такого, что $\langle \nabla f(x), h \rangle > 0$, верно $f^{\uparrow}_{Cl}(x, h) > 0$. Поэтому для некоторой последовательности $x'_k \to x$ $f(x'_k + \lambda h) - f(x'_k) > 0$ для достаточно малых $\lambda > 0$. Ввиду локальной липшицевости f в окрестности точки x это означает, что $f(x + \lambda h) - f(x) > 0$ для достаточно малых $\lambda > 0$ и выбранного направления h ($\langle \nabla f(x), h \rangle > 0$). Неравенство $f(y) - f(x) \geqslant 0$ следует теперь из непрерывности функционала f для всякого y такого, что $\langle \nabla f(x), y - x \rangle = 0$. Итак,

$$f(x) - f(x_*) \leqslant f(y) - f(x_*) \leqslant \omega(v_f(x, x_*)).$$

□

Отличительной особенностью данного утверждения является то, что мы рассматриваем не выпуклый, а квазивыпуклый целевой функционал f. Предположение о его локальной липшицевости позволяет в качестве аппарата для исследования дифференциальных свойств использовать субдифференциал Кларка.

Замечание 5.3.4. Обратим внимание [176], что для квазивыпуклого целевого функционала f и ограничения g вместо (суб)градиента $\nabla f(y)$ в $v_f(y, x_*)$ возможно рассмотреть некоторый элемент следующего множества

$$\widehat{D}f(x) = \{p \mid \langle p, x - y\rangle \geqslant 0 \ \ \forall y \in Q : \ f(y) < f(x)\}.$$

Следуя [139], предположим $\widehat{D}f(x) \neq \{0\}$ для $x \neq x_*$. Здесь и далее обозначим через $Df(x)$ один произвольный вектор из $\widehat{D}f(x)$:

$$Df(x) \in \widehat{D}f(x).$$

Такой подход может позволять работать и с квазивыпуклыми функционалами, для которых градиент (или субградиент) обращается в нуль в точках, отличных от глобального минимума. Дело в том, что вектор $Df(x)$ в таких случаях часто оказывается возможным выбрать ненулевым (см. [41, 42], а также пособие [16]).

На базе предыдущего утверждения и теоремы 5.3.2 можно оценить скорость сходимости алгоритма 17 для квазивыпуклого непрерывного целевого функционала f с L-липшицевым градиентом.

$$||\nabla f(x) - \nabla f(y)||_* \leqslant L||x - y|| \quad \forall x, y \in Q.$$

Применяя неравенство

$$f(x) \leqslant f(x_*) + ||\nabla f(x_*)||_* \cdot ||x - x_*|| + \frac{1}{2}L||x - x_*||^2,$$

получаем, что

$$\min_{k \in I} f(x^k) - f(x_*) \leqslant \min_{k \in I} \left\{ ||\nabla f(x_*)||_* \cdot ||x^k - x_*|| + \frac{1}{2}L||x^k - x_*||^2 \right\}.$$

Далее, справедливы оценки:

$$f(x) - f(x_*) \leqslant \varepsilon \cdot ||\nabla f(x_*)||_* + \frac{1}{2}L\varepsilon^2.$$

Поэтому справедливо

Следствие 5.3.5. *Пусть непрерывный квазивыпуклый функционал f имеет L-липшицев градиент ($L > 0$). Тогда после остановки алгоритма 17 справедлива оценка:*

$$\min_{1 \leqslant k \leqslant N} f(x^k) - f(x_*) \leqslant \varepsilon_f + \frac{L\varepsilon^2}{2}, \text{ где } \varepsilon_f = \varepsilon \cdot ||\nabla f(x_*)||_*.$$

Отметим, что по аналогии возможно выписывать оценки скорости сходимости для квазивыпуклых функционалов разного уровня гладкости по аналогии с выпуклым случаем (см. раздел 5.1). Основные результаты данного пункта показывают возможность сохранения полученных выше в выпуклом случае оценок и для алгоритма 20 для квазивыпуклых целевых функционалов соответствующего уровня гладкости (рассуждения полностью аналогичны проведённым для алгоритма 17).

5.4 Оптимальные адаптивные методы зеркального спуска для специальных типов негладких сильно выпуклых задач с функциональными ограничениями

В этом разделе работы мы рассмотрим задачу

$$f(x) \to \min, \quad g(x) \leqslant 0, \quad x \in Q \tag{5.23}$$

в предположениях сильной выпуклости f и g с одинаковым параметром $\mu > 0$. Слегка модифицируем предположения на прокс-функцию и допустим, что $d(x)$ ограничена на единичном шаре относительно выбранной нормы $\|\cdot\|$:

$$d(x) \leqslant \Theta_0^2, \quad \forall x \in Q : \|x\| \leqslant 1.$$

Также предположим, что для начальной точки $x^0 \in Q$ существует такое $R_0 > 0$, что $\|x_0 - x_*\|^2 \leqslant R_0^2$. Будем рассматривать методы для нахождения *ε-решения* $\widehat{x}$ поставленной задачи (5.23):

$$f(\widehat{x}) - f(x_*) \leqslant \varepsilon \text{ и } g(\widehat{x}) \leqslant \varepsilon.$$

Для построения методов решения задачи (5.23) при заданных предположениях мы используем идею рестартов (перезапусков) алгоритма 17 и алгоритма 18. Рассмотрим вспомогательное утверждение (см., например [11]).

Лемма 5.4.1. *Если f и g — μ-сильно выпуклые функционалы относительно нормы $\|\cdot\|$ на Q, $x_* = \arg\min\limits_{x\in Q} f(x)$, $g(x) \leqslant 0$ $(\forall x \in Q)$ и для некоторых $\varepsilon_f > 0$, а также $\varepsilon_g > 0$ верно:*

$$f(x) - f(x_*) \leqslant \varepsilon_f, \quad g(x) \leqslant \varepsilon_g.$$

Тогда

$$\frac{\mu}{2}\|x - x_*\|^2 \leqslant \max\{\varepsilon_f, \varepsilon_g\}.$$

Для определённости далее в этом разделе будем предполагать, что

$$f(x) = \max_{i=\overline{1,m}} f_i(x), \tag{5.24}$$

где f_i дифференцируемы во всякой точке $x \in Q$ и имеют липшицевы градиенты, т. е. существуют $L_i > 0$ такие, что

$$\|\nabla f_i(x) - \nabla f_i(y)\|_* \leqslant L_i\|x - y\| \quad \forall x, y \in Q.$$

Рассмотрим функцию $\tau : \mathbb{R}^+ \to \mathbb{R}^+$:

$$\tau(\delta) = \max\left\{\delta\|\nabla f(x_*)\|_* + \frac{\delta^2 L}{2},\ \delta\right\},$$

где $L := \max\limits_{i=\overline{1,m}}\{L_i\}$. Ясно, что функция τ возрастает и $\tau(0) = 0$, и поэтому для всякого $\varepsilon > 0$ существует $\widehat{\varphi}(\varepsilon) > 0 :\ \ \tau(\widehat{\varphi}(\varepsilon)) = \varepsilon$.

В [59] предложен следующий адаптивный алгоритм 23 для задачи (5.23).

Алгоритм 23 Адаптивный алгоритм зеркального спуска для сильно выпуклых функционалов.

Require: точность $\varepsilon > 0$; начальная точка x_0;
Θ_0 s.t. $d(x) \leqslant \Theta_0^2\ \forall x \in Q : \|x\| \leqslant 1$; Q; $d(\cdot)$;
параметр сильной выпуклости μ; R_0 удовлетворяет $\|x^0 - x_*\|^2 \leqslant R_0^2$.

1: Set $d_0(x) = d\left(\frac{x-x^0}{R_0}\right)$.
2: Set $p = 1$.
3: **repeat**
4: Set $R_p^2 = R_0^2 \cdot 2^{-p}$.
5: Set $\varepsilon_p = \frac{\mu R_p^2}{2}$.
6: Set x^p — выход алгоритма 17 с точностью $\widehat{\varphi}(\varepsilon_p)$, прокс-функцией $d_{p-1}(\cdot)$ и Θ_0^2.
7: $d_p(x) \leftarrow d\left(\frac{x-x^p}{R_p}\right)$.
8: Set $p = p + 1$.
9: **until** $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$.

Ensure: x^p.

Замечание 5.4.2. Возможно в пункте 6 листинга предыдущего метода просто использовать критерий остановки алгоритма 17. Тогда в итоговой оценке качества найденного решения по функции вместо ε будет $\tau(\varepsilon)$.

Замечание 5.4.3. В данном разделе все оценки скорости сходимости выписаны для целевых функционалов вида (5.24). Однако предлагаемый метод применим на всех классах μ-сильно выпуклых целевых функционалов, для которых рестартуемый алгоритм 17 приводит к оптимальным для выпуклых задач соответствующего уровня гладкости оценкам скорости сходимости. Только при этом в зависимости от выбранного класса функционалов необходимо соответствующим образом подбирать τ и $\widehat{\varphi}$. Например, для ν-гёльдеровых целевых функционалов ($\nu \in [0; 1]$) согласно замечанию 5.1.9 возможно выбрать $\tau(\delta) = \widehat{M}\delta$ (где $\widehat{M}$ удовлетворяет (5.15)).

Теорема 5.4.4. *Пусть f имеет вид* (5.24), *f и g — μ-сильно выпуклые функционалы на $Q \subset \mathbb{R}^n$ и $d(x) \leqslant \Theta_0^2$ для всех $x \in Q$ таких, что $\|x\| \leqslant 1$. Предположим, что начальное приближение $x^0 \in Q$ и число*

$R_0 > 0$ заданы так, что

$$\|x^0 - x_*\|^2 \leqslant R_0^2.$$

Тогда для $\widehat{p} = \left\lceil \log_2 \dfrac{\mu R_0^2}{2\varepsilon} \right\rceil$ *при достаточно большом* $M_g > 0$ *выход* $x_{\widehat{p}}$ *есть* ε*-решение задачи* (5.23)*, а также верны неравенства*

$$f(x^{\widehat{p}}) - f(x_*) \leqslant \varepsilon,\ g(x^{\widehat{p}}) \leqslant \varepsilon,\ \|x^{\widehat{p}} - x_*\|^2 \leqslant \frac{2\varepsilon}{\mu}.$$

При этом количество итераций алгоритма 17 при работе алгоритма 23 согласно пункту 6 листинга алгоритма 17 количество итераций алгоритма не превышает

$$\widehat{p} + \sum_{p=1}^{\widehat{p}} \frac{2\Theta_0^2 \max\{1, M_g^2\}}{\widehat{\varphi}^2(\varepsilon_p)}, \quad \text{где } \varepsilon_p = \frac{\mu R_0^2}{2^{p+1}}.$$

Доказательство. Функция $d_p(x) = d\left(\dfrac{x - x^p}{R_p}\right)$, определенная в алгоритме 23, является 1-сильно выпуклой функцией относительно нормы $\dfrac{\|.\|}{R_p}$ при всех $p \geqslant 0$. Методом математической индукции можно доказать, что

$$\|x^p - x_*\|^2 \leqslant R_p^2 \quad \forall p \geqslant 0.$$

Для $p = 0$ это утверждение очевидно ввиду выбора x^0 и R_0. Предположим, что для некоторого p верно $\|x^p - x_*\|^2 \leqslant R_p^2$. Докажем, что $\|x^{p+1} - x_*\|^2 \leqslant R_{p+1}^2$. Поскольку ввиду индуктивного допущения $d_p(x_*) \leqslant \Theta_0^2$, то согласно теореме 5.1.1 на $(p+1)$-м рестарте после не более чем (здесь при необходимости заменяем M_g на $R_0 M_g$)

$$N_{p+1} = \left\lceil \frac{2\Theta_0^2 \max\{1, M_g^2\}}{\widehat{\varphi}^2(\varepsilon_{p+1})} \right\rceil$$

итераций алгоритма 17, для $x^{p+1} = \bar{x}^{N_{p+1}}$ верны следующие неравенства:

$$f(x^{p+1}) - f(x_*) \leqslant \varepsilon_{p+1}, \quad g(x^{p+1}) \leqslant \varepsilon_{p+1} \quad \text{при} \quad \varepsilon_{p+1} = \frac{\mu R_{p+1}^2}{2}.$$

Тогда по лемме 5.4.1

$$\|x^{p+1} - x_*\|^2 \leqslant \frac{2\varepsilon_{p+1}}{\mu} = R_{p+1}^2.$$

Итак, для всякого $p \geqslant 0$ мы доказали, что

$$\|x^p - x_*\|^2 \leqslant R_p^2 = \frac{R_0^2}{2^p}, \quad f(x^p) - f(x_*) \leqslant \frac{\mu R_0^2}{2^{p+1}}, \quad g(x^p) \leqslant \frac{\mu R_0^2}{2^{p+1}}.$$

Поэтому при $p = \widehat{p} = \left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil$ выход x_p есть ε-решение задачи (5.23) и справедливы следующие неравенства

$$\|x^p - x_*\|^2 \leqslant R_p^2 = \frac{R_0^2}{2^p} \leqslant \frac{2\varepsilon}{\mu}.$$

Пусть K — общее число итераций алгоритма 17 при работе алгоритма 23 согласно пункту 6 листинга, а N_p — общее количество итераций алгоритма 17 на p-м рестарте. Вспомним, что функция $\tau : \mathbb{R}^+ \to \mathbb{R}^+$, возрастает и для каждого $\varepsilon > 0$ существует $\widehat{\varphi}(\varepsilon) > 0 : \tau(\widehat{\varphi}(\varepsilon)) = \varepsilon$. Поэтому имеем:

$$K = \sum_{p=1}^{\widehat{p}} N_p = \sum_{p=1}^{\widehat{p}} \left\lceil \frac{2\Theta_0^2 \max\{1, M_g^2\}}{\widehat{\varphi}^2(\varepsilon_p)} \right\rceil \leqslant \widehat{p} + \sum_{p=1}^{\widehat{p}} \frac{2\Theta_0^2 \max\{1, M_g^2\}}{\widehat{\varphi}^2(\varepsilon_p)}.$$

□

Замечание 5.4.5. Предыдущую оценку количества итераций работы алгоритма 17 можно несколько конкретизировать в случае $\varepsilon < 1$. В этом случае при всяком $\delta < 1$ имеем $\tau(\delta) \leqslant C\delta$ для некоторой константы $C > 0$. Поэтому можно считать, что $\widehat{\varphi}(\varepsilon) = \widehat{C} \cdot \varepsilon$ для соответствующей константы $\widehat{C} > 0$. Это означает, что для $\Omega = 2\Theta_0^2$ и при соответствующем выборе $\widehat{C} > 0$ ($\widehat{C} = \max\left\{1, R_p^{-1}\right\}$[1)], где $R_p^{-1} \geqslant 1$) на $p+1$-м перезапуске (рестарте) алгоритма 17 после не более, чем

$$k_{p+1} = \left\lceil \frac{\Omega \max\{1, M_g^2\} R_p^2}{\varepsilon_{p+1}^2} \right\rceil \tag{5.25}$$

итераций работы алгоритма 17, точка выхода x_{p+1} гарантированно удовлетворяет неравенству

$$f(x^{p+1}) - f(x_*) \leqslant \widehat{C} \cdot \varepsilon_{p+1}, \quad g(x^{p+1}) \leqslant \varepsilon_{p+1},$$

[1)] Указанная методика интересна, в первую очередь, при $R_p > 1$ по следующим причинам. Во-первых, при $R_p < 1$ и так имеется неплохое качество приближённого решения x^p по аргументу. Во-вторых, раз имеется столь удачная оценка качества исходной точки, то можно локализовать допустимое множество и задачи, на котором целевой функционал липшицев и применять уже известные методы (см. [68] и указанные там ссылки).

где $\varepsilon_{p+1} = \frac{\mu R_{p+1}^2}{2}$. Тогда по лемме 5.4.1,

$$\|x^{p+1} - x_*\|^2 \leqslant \frac{2\max\{1, \widehat{C}\}\varepsilon_{p+1}}{\mu} = \max\{1, \widehat{C}\} \cdot R_{p+1}^2.$$

Таким образом, для всякого $p \geqslant 0$,

$$\|x^p - x_*\|^2 \leqslant \max\{1, \widehat{C}\} \cdot R_p^2 = \max\{1, \widehat{C}\} \cdot R_0^2 \cdot 2^{-p}.$$

В то же время для произвольного $p \geqslant 1$ верны неравенства:

$$f(x^p) - f(x_*) \leqslant \max\{1, \widehat{C}\} \cdot \frac{\mu R_0^2}{2} \cdot 2^{-p}, \quad g(x_p) \leqslant \max\{1, \widehat{C}\} \cdot \frac{\mu R_0^2}{2} \cdot 2^{-p}.$$

Таким образом, если $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$, то x_p будет $\left(\max\{1, \widehat{C}\}\varepsilon\right)$-решением рассматриваемой задачи, причём:

$$\|x^p - x_*\|^2 \leqslant \max\{1, \widehat{C}\} \cdot R_0^2 \cdot 2^{-p} \leqslant \frac{2\varepsilon}{\mu},$$

если подобрать достаточно малое $\mu > 0$. Отметим, что если $\widehat{C} > 1$, то $R_{p_0} < 1$ и $\|x^{p_0} - x_*\|^2 < 1$.

Оценим теперь общее число N итераций алгоритма 17 согласно пункту 6 листинга алгоритма 23. Пусть $\widehat{p} = \left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil$. Тогда согласно (5.25), имеем с точностью до умножения на константу:

$$N = \sum_{p=1}^{\widehat{p}} k_p \leqslant$$

$$\leqslant \sum_{p=1}^{\widehat{p}} \left(1 + \frac{2\Theta_0^2 \max\{1, M_g^2\} R_p^2}{\varepsilon_{p+1}^2}\right) = \sum_{p=1}^{\widehat{p}} \left(1 + \frac{32\Theta_0^2 \max\{1, M_g^2\} 2^p}{\mu^2 R_0^2}\right) \leqslant$$

$$\leqslant \widehat{p} + \frac{64\Theta_0^2 \max\{1, M_g^2\} 2^{\widehat{p}}}{\mu^2 R_0^2} \leqslant \widehat{p} + \frac{64\Theta_0^2 \max\{1, M_g^2\}}{\mu\varepsilon}.$$

Теперь рассмотрим схему построения аналога алгоритма 20 [176] для сильно выпуклых задач. Это интересно, т.к. оценка качества решения по функционалу ограничения будет отличной от предыдущего метода рестартов алгоритма 17. Важно, что для алгоритма 20 с фиксированным количеством шагов (как и для алгоритма 17) можно получать оптимальные оценки скорости сходимости для целевых функционалов различного уровня гладкости [176].

Алгоритм 24 Адаптивный алгоритм зеркального спуска для сильно выпуклых функционалов, рестарты алгоритма 20.

Require: точность $\varepsilon > 0$; начальная точка x_0;
Ω s.t. $d(x) \leqslant \Omega^2\ \forall x \in Q : \|x\| \leqslant 1$; Q; $d(\cdot)$;
параметр сильной выпуклости μ; R_0 удовлетворяет $\|x^0 - x_*\|^2 \leqslant R_0^2$.

1: Set $d_0(x) = d\left(\frac{x-x^0}{R_0}\right)$.
2: Set $p = 1$.
3: **repeat**
4: Set $R_p^2 = R_0^2 \cdot 2^{-p}$.
5: Set $\varepsilon_p = \frac{\mu R_p^2}{2}$.
6: Set x^p — выход алгоритма 20 с точностью $\widehat{\varphi}(\varepsilon_p)$, прокс-функцией $d_{p-1}(\cdot)$ и Ω^2.
7: $d_p(x) \leftarrow d\left(\frac{x-x^p}{R_p}\right)$.
8: Set $p = p + 1$.
9: **until** $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$.

Теорема 5.4.6. *Пусть f имеет вид (5.24), а также f и g — μ-сильно выпуклые функционалы на $Q \subset \mathbb{R}^n$ и $d(x) \leqslant \Omega^2$ для всех $x \in Q$, таких, что $\|x\| \leq 1$, а также $M_g \leqslant 1$. Предположим, что начальное приближение $x^0 \in Q$ и число $R_0 > 0$ заданы так, что*

$$\|x^0 - x_*\|^2 \leqslant R_0^2.$$

Тогда для $\widehat{p} = \left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil$ выход $x^{\widehat{p}}$ есть ε-решение задачи (5.23), а также верны неравенства

$$f(x^{\widehat{p}}) - f(x_*) \leqslant \varepsilon, \quad g(x^{\widehat{p}}) \leqslant M_g \varepsilon,$$

$$\|x^{\widehat{p}} - x_*\|^2 \leqslant \frac{2\varepsilon}{\mu}.$$

При этом количество итераций алгоритма 20 при работе алгоритма 24 согласно пункту 6 листинга количество итераций алгоритма не превышает

$$\widehat{p} + \sum_{p=1}^{\widehat{p}} \frac{2\Omega^2}{\widehat{\varphi}^2(\varepsilon_p)}, \quad \text{где } \varepsilon_p = \frac{\mu R_0^2}{2^{p+1}}.$$

Доказательство. Функция $d_p(x) = d\left(\dfrac{x - x^p}{R_p}\right)$, определенная в алгоритме 24, является 1-сильно выпуклой функцией относительно нормы $\dfrac{\|\cdot\|}{R_p}$ при всех $p \geqslant 0$. Методом математической индукции можно доказать, что

$$\|x^p - x_*\|^2 \leqslant R_p^2 \quad \forall p \geqslant 0.$$

Для $p = 0$ это утверждение очевидно ввиду выбора x^0 и R_0. Предположим, что для некоторого p верно $\|x^p - x_*\|^2 \leqslant R_p^2$. Докажем, что $\|x^{p+1} - x_*\|^2 \leqslant R_{p+1}^2$. Поскольку ввиду индуктивного допущения $d_p(x_*) \leqslant \Omega^2$, то согласно теореме 5.1.2 на $(p+1)$-м рестарте после не более чем

$$N_{p+1} = \left\lceil \frac{2\Omega^2}{\widehat{\varphi}^2(\varepsilon_{p+1})} \right\rceil$$

итераций алгоритма 20, для $x^{p+1} = \bar{x}^{N_{p+1}}$ верны следующие неравенства:

$$f(x^{p+1}) - f(x_*) \leqslant \varepsilon_{p+1}, \quad g(x^{p+1}) \leqslant M_g \varepsilon_{p+1} \leqslant \varepsilon_{p+1} \quad \text{при} \quad \varepsilon_{p+1} = \frac{\mu R_{p+1}^2}{2}.$$

Тогда по лемме 5.4.1

$$\|x^{p+1} - x_*\|^2 \leqslant \frac{2\varepsilon_{p+1}}{\mu} = R_{p+1}^2.$$

Итак, для всякого $p \geqslant 0$ доказано, что

$$\|x^p - x_*\|^2 \leqslant R_p^2 = \frac{R_0^2}{2^p}, \quad f(x^p) - f(x_*) \leqslant \frac{\mu R_0^2}{2^{p+1}}, \quad g(x^p) \leqslant \frac{M_g \mu R_0^2}{2^{p+1}}.$$

Поэтому при $p = \widehat{p} = \left\lceil \log_2 \dfrac{\mu R_0^2}{2\varepsilon} \right\rceil$ выход x^p есть ε-решение задачи (5.23) и справедливы следующие неравенства

$$\|x^p - x_*\|^2 \leqslant R_p^2 = \frac{R_0^2}{2^p} \leqslant \frac{2\varepsilon}{\mu}.$$

Пусть K — общее число итераций алгоритма 20 при работе алгоритма 24 согласно пункту 6 листинга, а N_p — общее количество итераций алгоритма 20 на p-м рестарте. Вспомним, что функция $\tau : \mathbb{R}^+ \to \mathbb{R}^+$, возрастает и для каждого $\varepsilon > 0$ существует $\widehat{\varphi}(\varepsilon) > 0 : \tau(\widehat{\varphi}(\varepsilon)) = \varepsilon$. Поэтому имеем:

$$K = \sum_{p=1}^{\widehat{p}} N_p = \sum_{p=1}^{\widehat{p}} \left\lceil \frac{2\Omega^2}{\widehat{\varphi}^2(\varepsilon_p)} \right\rceil \leqslant \widehat{p} + \sum_{p=1}^{\widehat{p}} \frac{2\Omega^2}{\widehat{\varphi}^2(\varepsilon_p)}.$$

□

Замечание 5.4.7. Предыдущую оценку количества итераций работы алгоритма 20 можно несколько конкретизировать в случае $\varepsilon < 1$. В этом случае при всяком $\delta < 1$ имеем $\tau(\delta) \leqslant C\delta$ для некоторой константы $C > 0$. Поэтому можно считать, что $\widehat{\varphi}(\varepsilon) = \widehat{C} \cdot \varepsilon$ для соответствующей константы $\widehat{C} > 0$. Это означает, что (при соответствующем выборе $C > 0$, $\widehat{C} = \max\left\{1, R_{p_0}^{-1}\right\}$, где $R_{p_0}^{-1} \geqslant 1$) на $p+1$-м перезапуске алгоритма 20 после не более, чем

$$k_{p+1} = \left\lceil \frac{2\Omega^2 R_p^2}{\varepsilon_{p+1}^2} \right\rceil$$

итераций работы алгоритма 20, выход x^{p+1} гарантированно удовлетворяет неравенству

$$f(x^{p+1}) - f(x_*) \leqslant \widehat{C} \cdot \varepsilon_{p+1}, \quad g(x^{p+1}) \leqslant \varepsilon_{p+1},$$

где $\varepsilon_{p+1} = \frac{\mu R_{p+1}^2}{2}$. Тогда

$$\|x^{p+1} - x_*\|^2 \leqslant \frac{2\max\{1, \widehat{C}\}\varepsilon_{p+1}}{\mu} = \max\{1, \widehat{C}\} \cdot R_{p+1}^2.$$

Таким образом, всех $p \geqslant 0$,

$$\|x^p - x_*\|^2 \leqslant \max\{1, \widehat{C}\} \cdot R_p^2 = \max\{1, \widehat{C}\} \cdot R_0^2 \cdot 2^{-p}.$$

В то же время для всяких $p \geqslant 1$ верны неравенства:

$$f(x^p) - f(x_*) \leqslant \max\{1, \widehat{C}\} \cdot \frac{\mu R_0^2}{2} \cdot 2^{-p}, \quad g(x_p) \leqslant \max\{1, \widehat{C}\} \cdot \frac{\mu R_0^2}{2} \cdot 2^{-p}.$$

Таким образом, если $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$, то x^p будет $\left(\max\{1, \widehat{C}\}\varepsilon\right)$-решением рассматриваемой задачи, причём:

$$\|x^p - x_*\|^2 \leqslant \max\{1, \widehat{C}\} \cdot R_0^2 \cdot 2^{-p} \leqslant \frac{2\varepsilon}{\mu},$$

если подобрать достаточно малое $\mu > 0$.

Оценим теперь общее число N итераций алгоритма 20. Пусть $\widehat{p} = \left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil$. Тогда согласно (5.4.7), мы имеем с точностью до умножения на константу:

$$N = \sum_{p=1}^{\widehat{p}} k_p \leqslant \sum_{p=1}^{\widehat{p}} \left(1 + \frac{2\Omega^2 R_p^2}{\varepsilon_{p+1}^2}\right) = \sum_{p=1}^{\widehat{p}} \left(1 + \frac{32\Omega^2 2^p}{\mu^2 R_0^2}\right) \leqslant$$

$$\leqslant \widehat{p} + \frac{64\Omega^2 2^{\widehat{p}}}{\mu^2 R_0^2} \leqslant \widehat{p} + \frac{64\Omega^2}{\mu\varepsilon}.$$

Замечание 5.4.8. Результаты об оценке скорости сходимости предложенного метода сформулирован для класса μ-сильно выпуклых целевых функционалов вида (5.24) с μ-сильно выпуклыми липшицевыми функциональными ограничениями. По сути, для всех методов показано, что на таком классе ε-решение поставленной задачи по функции возможно достичь за $O(\varepsilon^{-1})$ итераций. Однако заметим, что можно гарантировать оптимальные оценки сложности (скорости сходимости) на классе μ-сильно выпуклых целевых функционалов, для которого удаётся установить оптимальные оценки скорости сходимости $O(\varepsilon^{-2})$ рестартуемого алгоритма 20 в выпуклом случае. Например, это верно на классе ν-гёльдеровых целевых функционалов ($\nu \in [0; 1]$) (см. замечание 5.4.3).

5.5 Адаптивный зеркальный спуск для задач онлайн оптимизации

Онлайн выпуклая оптимизация играет ключевую роль в решении задач, предполагающих обновление статистической информации [105,106]. Такого типа задачи возникают в разных прикладных задачах, связанных с моделированием интернет-сетей, наборов данных потребителей или финансового рынка. Задачи онлайн-оптимизации возникают, например, в машинном обучении [110]. Также важным примером является задача принятия решений [106, 113]. Предположим, нам дано N экспертов, а диапазон допустимых решений лежит на единичном симплексе. Каждый эксперт дает свои оценки потерь с возможным решением, и задача состоит в том, чтобы минимизировать общие потери с точки зрения всех экспертов (среднее арифметическое). Поэтому в последние годы активно развиваются методы решения задач онлайн-оптимизации [20, 21, 81, 82, 105, 106, 110, 126].

Предположим, что задано конечное число выпуклых липшицевых, вообще говоря, негладких функционалов на замкнутом подмножестве n-мерного векторного пространства. Задача онлайн-оптимизации заключается в том, чтобы минимизировать среднее арифметическое этих функционалов при наличии выпуклого липшицева негладкого ограни-

чения. Точнее говоря, пусть заданы N выпуклых липшицевых негладких функционалов $f_i(x)$ заданы на выпуклом замкнутом подмножестве n-мерного векторного пространства. Рассмотрим задачу минимизации среднего арифметического нескольких функционалов с одним выпуклым липшицевым ограничением

$$\frac{1}{N}\sum_{i=1}^{N} f_i(x) \to \min_{x\in Q},\ g(x) \leqslant 0 \tag{5.26}$$

Важной особенностью является возможность вычислять (суб)градиент $\nabla f_i(x)$ всего один раз.

В [180], в частности, предложен адаптивный алгоритм зеркального спуска, который гарантирует достижение приемлемого приближённого решения с заданной точностью с малым количество шагов непродуктивных шагов (не более, чем $O(N)$). Отличительная особенность наших подходов по сравнению с известными аналогами — возможность использования неевклидовой прокс-структуры, что существенно для некоторых задач онлайн-оптимизации (например, задача об экспертах приводит к задаче на единичном симплексе). После остановки методов [180] выполнено неравенство:

$$\frac{1}{N}\sum_{i=0}^{N} f_i(x^k) - \min_{x\in Q}\frac{1}{N}\sum_{i=0}^{N} f(x) \leqslant \delta.$$

Показано, что если левая часть неравенства предыдущего (эту величину называют регретом) неотрицательна, а точность δ удовлетворяет $\delta \leqslant \varepsilon = \frac{C}{\sqrt{N}}$ для некоторой константы $C > 0$, то непродуктивных шагов будет не более, чем $O(N)$. В [180] показано, что в случае, когда регрет отрицателен, непродуктивных шагов будет не более, чем $O(N^2)$, однако может быть достигнута гораздо большая точность решения задачи. Отметим, что в [180] также исследована модификация предложенного алгоритма для случая большого числа ограничений и приводятся примеры расчётов для демонстрации их преимуществ.

Предположим, что выполняется N продуктивных шагов, и на каждом шаге вычисляется (суб) градиент ровно одного целевого функционала. Обозначим число непродуктивных шагов как N_J.

Теперь рассмотрим адаптивный метод (алгоритм 25) для задач (5.26). Основной особенностью по сравнению с другими методами из

Алгоритм 25 Задачи онлайн-оптимизация с функциональными ограничениями: адаптивный зеркальный спуск

Require: $\varepsilon, N, \Theta_0^2, Q, d(\cdot), x^0$

$i := 1,\ k := 0;$
repeat
  if $g(x^k) \leqslant \varepsilon$ **then**
    $M_k := \|\nabla f_i(x^k)\|_*;$
    $h_k = \Theta_0 \left(\sum\limits_{t=0}^{k} M_t^2 \right)^{-1/2};$
    $x^{k+1} := Mirr[x^k](h_k \nabla f_i(x^k));$
    $i := i + 1;$
    $k := k + 1;$
  else
    $M_k := \|\nabla g(x^k)\|_*;$
    $h_k = \Theta_0 \left(\sum\limits_{t=0}^{k} M_t^2 \right)^{-1/2};$
    $x^{k+1} := Mirr[x^k](h_k \nabla g(x^k));$
    $k := k + 1;$
  end if
until $i = N + 1$
Guaranteed accuracy:

$$\delta := \frac{2\Theta_0}{N} \left(\sum_{i=0}^{N+N_J-1} M_i^2 \right)^{1/2} - \varepsilon \cdot \frac{N_J}{N}. \tag{5.27}$$

настоящего раздела является неубывающий размер шага с учетом нормы (суб) градиента целевой функции или ограничений на конкретной итерации. Поэтому предложенный алгоритм будет работать до тех пор, пока не получится ровно N продуктивных шагов. В результате мы получаем последовательность $\{x^k\}_{k \in I}$ на продуктивных шагах, которую можно рассматривать как решение задачи (5.26) с точностью δ.

Согласно лемме 5.0.1

$$f_i(x^k) - f_i(x) \leqslant \frac{h_k}{2} \|\nabla f_i(x^k)\|_*^2 + \frac{V(x, x^k)}{h_k} - \frac{V(x, x^{k+1})}{h_k}$$

$$g(x^k) - g(x) \leqslant \frac{h_k}{2}\|\nabla g(x^k)\|_*^2 + \frac{V(x,x^k)}{h_k} - \frac{V(x,x^{k+1})}{h_k}$$

Выполнив деление каждого неравенства на h_k и просуммировав их по k от 0 до $N+N_J-1$ с учётом выражения для шагов h_k, получаем

$$\sum_{k\in I}\big(f(x^k)-f(x_*)\big)+\sum_{k\in J}\big(g(x^k)-g(x_*)\big)\leqslant \sum_{k=0}^{N+N_J-1}\frac{h_k M_k^2}{2}+$$

$$+\sum_{k=0}^{N+N_J-1}\frac{1}{h_k}\big(V(x_*,x^k)-V(x_*,x^{k+1})\big)\ \text{и}$$

$$\sum_{k=0}^{N+N_J-1}\frac{1}{h_k}\big(V(x_*,x^k)-V(x_*,x^{k+1})\big)=$$

$$=\frac{1}{h_0}V(x_*,x^0)+\sum_{k=0}^{N+N_J-2}\Big(\frac{1}{h_{k+1}}-\frac{1}{h_k}\Big)V(x_*,x^{k+1})-$$

$$-\frac{1}{h_{N+N_J-1}}V(x_*,x^k)\leqslant\frac{\Theta_0^2}{h_0}+\Theta_0^2\sum_{k=0}^{N+N_J-2}\Big(\frac{1}{h_{k+1}}-\frac{1}{h_k}\Big)=\frac{\Theta_0^2}{h_{N+N_J-1}}.$$

Поэтому по определению размеров шагов h_k,

$$\sum_{i=1}^{N}\big(f_i(x^k)-f(x_*)\big)+\sum_{k\in J}\big(g(x^k)-g(x_*)\big)\leqslant$$

$$\leqslant\sum_{k=0}^{N+N_J-1}\frac{h_k M_k^2}{2}+\frac{\Theta_0^2}{h_{N+N_J-1}}\leqslant$$

$$\leqslant\sum_{k=0}^{N+N_J-1}\frac{\Theta_0}{2}\frac{M_k^2}{\left(\sum_{j=0}^{k}M_j^2\right)^{1/2}}+\Theta_0\left(\sum_{k=0}^{N+N_J-1}M_k^2\right)^{1/2}\leqslant$$

$$\leqslant 2\Theta_0\left(\sum_{k=0}^{N+N_J-1}M_k^2\right)^{1/2},$$

где мы использовали неравенство

$$\sum_{i=0}^{N+N_J-1}\frac{M_i^2}{\left(\sum_{j=0}^{i}M_j^2\right)^{1/2}}\leqslant 2\left(\sum_{i=0}^{N+N_J-1}M_i^2\right)^{1/2},$$

что можно доказать по индукции. Поскольку для $k \in J$, $g(x^k) - g(x_*) \geqslant g(x^k) > \varepsilon$, мы получаем

$$\sum_{i=1}^{N}(f_i(x^k) - f_i(x^*)) < \varepsilon N - \varepsilon(N + N_J) + 2\Theta_0 \left(\sum_{i=0}^{N+N_J-1} M_i^2 \right)^{1/2}. \quad (5.28)$$

и в силу (5.27)

$$\frac{1}{N}\sum_{i=1}^{N} f_i(x^k) - \min_{x \in Q} \frac{1}{N}\sum_{i=1}^{N} f_i(x) \leqslant \delta.$$

Если мы примем неотрицательность регрета (то есть левой части (5.28)) и точность будет определена согласно (5.27), то верно

$$\varepsilon(N + N_J) \leqslant \varepsilon N + 2\Theta_0 \left(\sum_{i=0}^{N+N_J-1} M_i^2 \right)^{1/2} \leqslant \varepsilon N + 2M\Theta_0 \cdot \sqrt{N + N_J},$$

$$N_J^2 \leqslant \frac{4M^2\Theta_0^2(N + N_J)}{\varepsilon^2} = \frac{4M^2\Theta_0^2(N + N_J)N}{C^2}$$

Далее,

$$\frac{N_J^2}{N^2 + NN_J} = \frac{\left(\frac{N_J}{N}\right)^2}{1 + \frac{N_J}{N}} \leqslant \frac{4M^2\Theta_0^2}{C^2}$$

и $N_J = O(N)$. Таким образом, справедлив следующий результат.

Теорема 5.5.1. *Предположим, что алгоритм 25 работает ровно N продуктивных шагов. Тогда после его остановки верно следующее неравенство:*

$$\frac{1}{N}\sum_{i=1}^{N} f_i(x^k) - \min_{x \in Q} \frac{1}{N}\sum_{i=1}^{N} f_i(x) \leqslant \delta.$$

Если предположить $\delta \leqslant \varepsilon = \frac{C}{\sqrt{N}}$, а также неотрицательность регрета

$$\frac{1}{N}\sum_{i=1}^{N} f_i(x^k) - \min_{x \in Q} \frac{1}{N}\sum_{i=1}^{N} f_i(x) \geqslant 0,$$

то количество непродуктивных шагов не превысит $O(N)$.

Замечание 5.5.2. Полученный результат в частности означает, что алгоритм 25 оптимален с точностью до константы для рассматриваемого класса задач онлайн-оптимизации [105].

5.6 О применимости разработанных адаптивных зеркальных спусков к некоторым прикладным задачам

5.6.1 Приложения к задаче оптимизации высоконагруженной компьютерной сети В данном пункте покажем, как предложенные методы зеркального спуска с переключением могут быть применены к задаче распределения ресурсов, в частности к задаче оптимизации компьютерной сети. Уточним постановку задачи [50, 109]. Допустим, что имеется компьютерная сеть с n пользователями (узлами), которые обмениваются пакетами через фиксированный набор m соединений. Структура сети задана матрицей маршрутизации $C = (C_i^j) \in \mathbb{R}^{m\times n}$, столбцы которой $\mathbf{C}_i \neq 0$, $i = 1, \ldots, n$ есть булевы m-мерные векторы такие, что $C_i^j = 1$ в случае использования узлом i соединения j, в противном случае $C_i^j = 0$. Ограничения на пропускную способность соединений задаются вектором $\mathbf{b} \in \mathbb{R}_+^m$ со строго положительными компонентами. Пользователи оценивают качество работы сети с помощью функций полезности $u_p(x_p)$, $p = 1, \ldots, n$, где $x_p \in \mathbb{R}_+^m$ – скорость передачи данных k–го пользователя. Согласно [117] в качестве критерия оптимальности системы можно принять сумму функций полезности для всех пользователей. В этом случае задачу максимизации суммарной полезности сети при заданных ограничениях на пропускную способность соединений можно сформулировать следующим образом:

$$\max_{\left\{C\mathbf{x}=\sum\limits_{p=1}^{n}\mathbf{C}_p x_p\right\}\leqslant \mathbf{b}} \left\{U(\mathbf{x}) = \sum_{p=1}^{n} u_p(x_p)\right\}, \tag{5.29}$$

где $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}_+^n$. Решением данной задачи будет оптимальное распределение ресурсов x_*. Отметим, что целевые функционалы тут могут быть самого разного уровня гладкости, в том числе даже формально не удовлетворяющие условию Липшица: $u_p(x) = \ln x$, $u_p(x) = \sqrt{x_p}$.

Обозначим $g_j(\mathbf{x}) = \langle \mathbf{C}_j, \mathbf{x}\rangle - b_j$, $j = 1, \ldots, m$. Ясно, что

$$|g_j(\mathbf{x}^1) - g_j(\mathbf{x}^2)| \leqslant \|\mathbf{C}_j\|_2 \|\mathbf{x}^1 - \mathbf{x}^2\|_2 \leqslant m \cdot \|\mathbf{x}^1 - \mathbf{x}^2\|_2.$$

Отметим, что последнее неравенство верно в силу задания матрицы C. Итак, всякий функционал $g_j(\mathbf{x})$ при $j = 1, \ldots, m$ $M_g = m$ удовлетворяет условию Липшица.

Рассмотрим задачу выпуклой минимизации, эквивалентную (5.29)

$$\min_{g_j(\mathbf{x})\leqslant 0,\, j=\overline{1,m}} f(\mathbf{x}),$$

где $f(\mathbf{x}) = -U(\mathbf{x})$.

Модификация зеркальных спусков для логарифмической функции полезности.

Отметим, что наиболее распространены логарифмические функции полезности, то есть $u_p(x_p) = \log x_p$. Такие функции не удовлетворяют свойству Липшица на $\mathbb{R}^n_+$, поскольку градиент неограничен вблизи нуля. Однако в такой ситуации можно рассмотреть следующую модификацию алгоритма 20. Отодвинем границу допустимого множества от нуля, то есть положим $x_k \geqslant \varepsilon n$, $k = 1, \ldots, n$. Тогда функция полезности будет липшицевой с постоянной $M_U = \frac{1}{\varepsilon}$, то есть

$$||\nabla U(\mathbf{x})||_2 \leqslant \sum_{p=1}^{n} |u_p'(x_p)| \leqslant n \cdot \frac{1}{\varepsilon n} = \frac{1}{\varepsilon}.$$

Это означает, что возможно применить алгоритм 20 для $N = \left\lceil 2\frac{\Theta_0^2}{\varepsilon^4} \right\rceil$ с шагом $h_k = \frac{\varepsilon^2}{\|\nabla f(\mathbf{x}^k)\|_2}$ для $k \in I$ и $h_k = \frac{\varepsilon^2}{\|\nabla g_{j_k}(\mathbf{x}^k)\|_2}$ для $k \notin I$. Из теоремы 5.2.4 получаем

Следствие 5.6.1. *После выполнения $N = \left\lceil 2\frac{\Theta_0^2}{\varepsilon^4} \right\rceil$ шагов алгоритма 20 выполняется неравенство $\min\limits_{k\in I} f(\mathbf{x}^k) - f(\mathbf{x}^*) \leqslant \varepsilon$.*

Оценим скорость сходимости алгоритма 19, примененного к этой задаче. Из теоремы 5.2.1 и замечания после нее получаем

Следствие 5.6.2. *После $N = \left\lceil 2\frac{\Theta_0^2}{\varepsilon^4} \right\rceil$ шагов алгоритма 19, выполняется неравенство $f(\widehat{\mathbf{x}}^N) - f(\mathbf{x}^*) \leqslant \varepsilon$, где $\widehat{\mathbf{x}}^N = \frac{1}{\sum_{k\in I} h_k} \sum\limits_{k\in I} h_k \mathbf{x}^k$.*

Обратим внимание, что оценки скорости сходимости алгоритма 20 и алгоритма 19 имеют один и тот же порядок. Однако благодаря адаптивности критерия остановки алгоритм 19 работает лучше на практике. Более того, алгоритм 19 не требует модификации шагов для отступов от нуля.

При этом отметим, что ввиду следствия 5.3.5 алгоритмы 17 и 20 приводят к оптимальным оценкам сложности для гельдеровых функций полезности. Для логарифмической функции полезности $u_p(x) = \ln x_p$ можно модифицировать допустимое множество задачи отступами от нуля по значениям переменных ($x_p \geqslant n\varepsilon$) и показать оценку сложности $O\left(\varepsilon^{-4}\right)$ для алгоритмов 17 и 20. Если же для некоторого $R > 0$ добавить условие ограниченности $\|x\|_2 \leqslant R$, то можно утверждать, что $\forall x, y \in Q$

$$|u_k(x) - u_k(y)| \leqslant 2 \max\left\{\ln R, |\ln n\varepsilon|\right\},$$

Тогда целевую функцию задачи можно считать гёльдеровой с показателем $\nu = 0$, причем $M_0 = 2n \max\left\{\ln R, |\ln n\varepsilon|\right\}$, откуда

$$f(\widehat{x}) - f(x_*) \leqslant 4n^2 \max\left\{\ln^2 R, \ln^2 n\varepsilon\right\} \cdot \varepsilon.$$

Это означает, что оценку сложности $O\left(\frac{1}{\varepsilon^4}\right)$ для алгоритмов 17 и 20 можно заменить на

$$O\left(\frac{n^4 \ln^4 n\varepsilon}{\varepsilon^2}\right),$$

поскольку вместо ε в оценке количества итераций (то есть (5.19)) достаточно выбрать $\frac{\varepsilon}{4n^2 \max\{\ln^2 R, \ln^2 n\varepsilon\}}$. Однако при этом оценка сложности зависит от величины, соответствующей количеству пользователей. Отметим при этом, что для проведённых экспериментов получилось, что алгоритм 19 за счёт адаптивного критерия остановки работает существенно быстрее по сравнению с алгоритмом 20. Опишем результаты указанных экспериментов в следующем примере.

Пример 5.6.3. Пусть дана матрица C размера 20×200, состоящая из нулей и единиц, и вектор $b \in \mathbb{R}^{200}$, состоящий из чисел, принадлежащих интервалу $[1, 5]$. Рассмотрим задачу максимизации суммарной полезности сети при заданных ограничениях (5.29) и с целевым функционалом $u_p(x) = \sqrt{x_p}$. Выбрана точка старта $x^0 = (0.001, \ldots, 0.001)$. Сравниваются количество итераций и время работы алгоритмов 19 и 20 для различных значений ε. Усредненные результаты пяти экспериментов представлены в таблице 5.1. Очевидно, целевой функционал данной задачи удовлетворяет условию Гёльдера. Это означает, что для алгоритма 20 мы можем доказать оптимальную оценку сложности $O(\varepsilon^{-2})$, для алгоритма 19 мы ее доказать не можем. Однако за счет адаптивности этого алгоритма мы получаем лучшее качество решения.

Таблица 5.1. Результаты выполнения алгоритмов 19 и 20, пример 5.6.3.

ε	Алгоритм 20		Алгоритм 19	
	Итерации	Время, с	Итерации	Время, с
$1/2$	40000	106	706.8	1
$1/3$	90000	234	1565.6	3
$1/4$	160000	425	2808.4	7
$1/5$	250000	649	4439.8	10
$1/6$	360000	925	6467	16

Таблица 5.2. Результаты выполнения алгоритмов 19 и 20, пример 5.6.4.

ε	Алгоритм 20		Алгоритм 19	
	Итерации	Время, с	Итерации	Время, с
$1/2$	40000	108	3814	9
$1/3$	90000	236	9341.8	23
$1/4$	160000	425	17553.2	45
$1/5$	250000	646	28488.4	73
$1/6$	360000	926	42187	110

Пример 5.6.4. Пусть дана матрица C размера 20×200, состоящая из нулей и единиц, и вектор $b \in \mathbb{R}^{200}$, состоящий из чисел, принадлежащих интервалу $[1, 5]$. Рассмотрим задачу максимизации суммарной полезности сети при заданных ограничениях (5.29) и с целевым функционалом $u_p(x) = \ln x_p$. Выбрана точка старта $x^0 = (0.001, \ldots, 0.001)$. Сравниваются количество итераций и время работы алгоритмов 19 и 20 для различных значений ε. Усредненные результаты пяти экспериментов представлены в таблице 5.2. В этом случае целевая функция не удовлетворяет даже условию Гёльдера. Однако за счет остановки адаптивного алгоритма мы можем гарантировать приемлемое качество решения.

Для проверки эффективности алгоритма 19 было проведено сравнение скорости его работы с методом эллипсоидов (алгоритм 26, [137]) для задачи, двойственной к указанной (переход к двойственной задаче целесообразен ввиду наличия аффинных функциональных ограничений). Этот метод выбран для сравнения, поскольку он сходится с линейной скоростью и в работе [137] детально теоретически проработан вопрос восстановления приближённого решения прямой задачи по найденному приближённому решению двойственной задачи. В [26, 109] (преимуще-

ственно нашими коллегами) получена следующая теоретическая оценка скорости сходимости метода эллипсоидов для указанной задачи (точнее говоря, для двойственной к ней задачи).

Алгоритм 26 Метод эллипсоидов

Require: $u_k(x_k), k = 1, ..., n$ — вогнутые функции.

1: $B_0 := 2R \cdot I_n$
2: **for** $t = 0, ..., N-1$ **do**
3: Вычислить $\nabla\varphi(\boldsymbol{\lambda}^t)$
4: $\mathbf{q}_t := B_t^T \nabla\varphi(\boldsymbol{\lambda}^t)$
5: $\mathbf{p}_t := \dfrac{B_t^T \mathbf{q}_t}{\sqrt{\mathbf{q}_t^T B_t B_t^T \mathbf{q}_t}}$
6:
$$B_{t+1} := \frac{m}{\sqrt{m^2-1}} B_t + \left(\frac{m}{m+1} - \frac{m}{\sqrt{m^2-1}} \right) B_t \mathbf{p}_t \mathbf{p}_t^T$$
7: $\boldsymbol{\lambda}^{t+1} := \boldsymbol{\lambda}^t - \dfrac{1}{m+1} B_t \mathbf{p}_t$
8: **end for**
9:
10: **return** $\boldsymbol{\lambda}^N$

Теорема 5.6.5. *Пусть начальная точка для алгоритма 26 лежит в единичном шаре $B_0 = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \|\boldsymbol{\lambda}\|_2 \leqslant 2R\}$. Тогда после*

$$N = 2m(m+1) \left\lceil \log \left(\frac{32 \cdot 4MR}{\varepsilon} \right) \right\rceil \tag{5.31}$$

итераций для некоторой точки $\widehat{x}^N$ будут верны следующие неравенства

$$U(\mathbf{x}^*) - U(\widehat{\mathbf{x}}^N) \leqslant \varepsilon, \ \|[C\widehat{\mathbf{x}}^N - \mathbf{b}]_+\|_2 \leqslant \varepsilon.$$

Таким образом, скорость сходимости согласно оценке (5.31) линейная. Однако при достаточно больших m ясно, что оценка (5.31) становится неприемлемой. Это можно показать на примере расчетов. Поведение методов было проверено экспериментально в задачах (5.29) различных конфигураций сетей и с разной точностью ε. Вычислительные эксперименты выполнены учащимся ФМЛ 239 г. Санкт-Петербурга Дмитрием Аркадьевичем Пасечнюком.

Матрица маршрутизации C сгенерирована следующим образом: $C_i^j = 1$ с вероятностью $p = 0.5$ или $C_i^j = 0$ с вероятностью $1 - p = 0.5$.

Элементы вектора $\mathbf{b}$ являются равномерными случайными величинами: $b_i \in [0.1, 0.4]$. Функции полезности являются логарифмическими. Начальные значения для алгоритма 19 и метода эллипсоидов (МЭ): $\mathbf{x}^0 = 0$ и $\boldsymbol{\lambda}^0 = 10^{-20}$ соответственно. Радиус $2R$ начального шара в методе эллипсоидов и радиус $R = \sqrt{2}\Theta$ шара, содержащего $\mathbf{x}_*$, в алгоритме 19 были выбраны экспериментально таким образом, чтобы промежуточные решения, полученные данными методами, оставались на каждой итерации внутри заданного множества. Требуемая точность решения $n\varepsilon$ была выбрана так, чтобы граничный сдвиг $n\varepsilon$ допустимого множества с нуля был достаточно мал, то есть не более $\sim 10^{-1}$.

Таблица 5.3. Результаты сходимости алгоритма 19 (А2) и метода эллипсоидов (МЭ),

$$\varepsilon = 6e - 4$$

	n	50		100		200	
	m	100	150	100	150	100	150
А2	Итерации	142243	142516	171292	174270	193621	198585
	Время, с	16.77	21.91	33.56	37.63	46.8	49.22
МЭ	Итерации	512749	758327	531448	760537	532992	761008
	Время, с	601.74	885.54	1022.73	1293.15	1418.67	1481.02

Таблица 5.4. Результаты сходимости алгоритма 19 (А2) и метода эллипсоидов (МЭ),

$$\varepsilon = 3e - 4$$

	n	50		100		200	
	m	100	150	100	150	100	150
А2	Итерации	8724510	9105234	9006192	9574296	9157003	9611472
	Время, с	921.38	1224.70	1276.73	1411.67	1424.12	1670.74
МЭ	Итерации	603578	801775	628267	833323	633571	850051
	Время, с	1084.22	1317.61	1321.48	1705.46	1492.88	1921.06

Результаты экспериментов представлены в таблицах 5.3–5.5. Как видно из таблиц 5.3–5.4, для $\varepsilon = 6e - 4$ и $\varepsilon = 3e - 4$ предлагаемый алгоритм показывает лучшее время, чем метод эллипсоидов. Даже для

Таблица 5.5. Результаты сходимости алгоритма 19 (А2) и метода эллипсоидов (МЭ),

$$\varepsilon = 2e - 4$$

	n	50		100	
	m	100	150	100	150
А2	Итерации	25225735	29752323	26055762	33846145
	Время, с	1367.54	1569.25	1550.62	1796.34
МЭ	Итерации	599423	960529	618783	971525
	Время, с	1223.86	1515.32	1677.25	1900.03
	n	200			
	m	100	150		
А2	Итерации	28532359	37244837		
	Время, с	1723.07	1985.52		
МЭ	Итерации	667294	1021528		
	Время, с	1891.18	2293.34		

высокой точности решения $\varepsilon = 2e - 4$ алгоритм 19 показал большое количество итераций и почти столько же времени, сколько МЭ, как показано в таблице 5.5. Таким образом, в случае, когда очень высокая точность решения не требуется, вполне разумно применить предложенный алгоритм.

Проведенные эксперименты подтверждают отмеченный выше теоретический факт об алгоритме 19: скорость сходимости алгоритма 19 зависит только от уровня гладкости целевой функции и ограничений и не зависит от числа ограничений m. В отличие от случая с методом эллипсоидов, где при теоретическом числе итераций квадратичный рост по m имеет квадратичный рост (теорема 5.6.5). Таким образом, если мы сравним количество итераций для тех же n и $m = 100$, $m = 150$ в таблицах 5.3–5.5. Можно заметить, что число итераций для метода эллипсоидов увеличивается почти в 1,5–2 раза, и это не так для алгоритма 19. Теоретические результаты говорят , что для алгоритма 19 (m увеличивается и n одинаково) число итераций не должно изменяться. Но из-за адаптивности критерия остановки на практике он изменяется незначительно, поскольку в обоих случаях это число меньше теоретической оценки скорости сходимости.

Таблица 5.6. Результаты выполнения алгоритмов 19, 20 и модификации алгоритма 20.

ε	Алгоритм 20		Алгоритм 20 (версия 2)		Алгоритм 19	
	Итерации	Время, с	Итерации	Время, с	Итерации	Время, с
$1/2$	62800.2	26.4	62800.2	26.6	1317	0.01
$1/3$	141300.2	60.6	141300.2	60.2	2895	1
$1/4$	251200.2	108	251200.2	106.2	5116.4	1.8
$1/5$	392500	167.2	392500	168	7943.6	3.2
$1/6$	565200.2	241	565200.2	250.6	11381	4.6

5.6.2 Применимость рассматриваемых методов к задаче проектирования механических конструкций (Truss Topology Design)

Рассмотрим результаты экспериментов для задачи типа (1.38) для алгоритмов 19 и 20.

Пример 5.6.6. Пусть дана матрица C размера 200×20, состоящая из чисел, принадлежащих интервалу $[-10, 10]$. Рассмотрим задачу выпуклой минимизации (1.38), вектор f состоит из чисел, принадлежащих интервалу $[1, 10]$. Выбрана точка старта $x^0 = (0.001, \ldots, 0.001)$. Сравниваются количество итераций и время работы алгоритмов 19, 20 и версия алгоритма 20 для различных значений ε. Усредненные результаты пяти экспериментов представлены в таблице 5.6. Отметим, что похожий на алгоритм 20 метод рассматривался в [119].

5.6.3 Численные эксперименты

Сравним скорость работы алгоритмов 17 и 20 для некоторых примеров задач минимизации при функциональных ограничениях.

Пример 5.6.7. Аналог задачи о наименьшем покрывающем шаре. Входные данные: $n = 1000$, координаты точек $A_k = (a_{1k}, a_{2k}, \ldots, a_{nk})$ $(k = 1, 2, \ldots, 5)$ — целые числа из интервала $[-10, 10]$, целевая функция $(M_f = 1)$

$$f(x) = \max_{k=\overline{1,5}} \left(\sqrt{(x_1 - a_{1k})^2 + (x_2 - a_{2k})^2 + \ldots + (x_n - a_{nk})^2} \right),$$

Таблица 5.7. Сравнение результатов работы алгоритмов, пример 5.6.7.

ε	Итерации	Время, с	Итерации	Время, с
	Алгоритм 17		Алгоритм 20	
$1/2$	—	>300	16	0.071
$1/4$	—	>300	64	0.259
$1/6$	—	>300	144	0.575
$1/8$	—	>300	256	1

$x^0 = \frac{(0.1,\ldots,0.1)}{\|(0.1,\ldots,0.1)\|}$, функционалы ограничений

$$\begin{aligned} g(x) &= \max_{m=1,2,3,\ldots,20}\{g_m(x)\} \leqslant 0, \\ g_1(x) &= \alpha_{11}|x_1| + \alpha_{12}|x_2| + \ldots + \alpha_{1n}|x_n| - 1, \\ g_2(x) &= \alpha_{21}|x_1| + \alpha_{22}|x_2| + \ldots + \alpha_{2n}|x_n| - 1, \\ &\ldots \\ g_m(x) &= \alpha_{m1}|x_1| + \alpha_{m2}|x_2| + \ldots + \alpha_{mn}|x_n| - 1. \end{aligned} \tag{5.32}$$

Коэффициенты $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{mn}$ представлены матрицей

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \ldots & 1 & 1 \\ 1 & 2 & 2 & 2 & \ldots & 2 & 2 \\ 1 & 3 & 3 & 3 & \ldots & 3 & 3 \\ 1 & 2 & 3 & 4 & \ldots & 999 & 1000 \\ 1 & 3 & 4 & 5 & \ldots & 1000 & 1001 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 18 & 19 & 20 & \ldots & 1015 & 1016 \end{pmatrix}. \tag{5.33}$$

Результаты для данного примера представлены в таблице 5.7. Заметим, что алгоритм 20 работает быстрее, чем алгоритм 17. При этом ввиду $M_f = 1$ выполнение критериев остановки алгоритмов 17 и 20 приводит к сопоставимому качеству решения задачи.

Пример 5.6.8. Аналог задачи Ферма–Торричелли–Штейнера. Входные данные: $n = 1000$, координаты точек $A_k = (a_{1k}, a_{2k}, \ldots, a_{nk})$ $(k = 1, 2, \ldots, 5)$ — целые числа из интервала $[-10, 10]$, целевая функция $(M_f = 1)$

$$f(x) = \frac{1}{5}\sum_{k=1}^{5}\sqrt{(x_1 - a_{1k})^2 + (x_2 - a_{2k})^2 + \ldots + (x_n - a_{nk})^2},$$

Таблица 5.8. Сравнение результатов работы алгоритмов, пример 5.6.8.

ε	Итерации	Время, с	Итерации	Время, с
	Алгоритм 17		Алгоритм 20	
$1/2$	—	>300	16	0.068
$1/4$	—	>300	64	0.264
$1/6$	—	>300	144	0.526
$1/8$	—	>300	256	0.920

$x^0 = \frac{(0.1,\ldots,0.1)}{\|(0.1,\ldots,0.1)\|}$, функционалы ограничений представлены (5.32). Коэффициенты $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{mn}$ представлены матрицей (5.33).

Результаты для данного примера представлены в таблице 5.8. Заметим, что алгоритм 20 работает быстрее, чем алгоритм 17. При этом ввиду $M_f = 1$ выполнение критериев остановки алгоритмов 17 и 20 приводит к сопоставимому качеству решения задачи.

Пример 5.6.9. Пример вогнутой целевой функции, удовлетворяющей условию Гёльдера. Пусть $n = 1000$, целевой функционал ($M_{f,1/2} = 1$)

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}\sqrt{x_i},$$

$x^0 = \frac{(0.1,\ldots,0.1)}{\|(0.1,\ldots,0.1)\|}$, $Q = \{x = (x_1,\ldots,x_n) \quad | \quad x_i \geqslant 0 \quad \forall i, \ \sum\limits_{i=1}^{n} x_i^2 \leqslant 1\}$. Функционалы ограничений имеют вид

$$\begin{aligned} g(x) &= \max_{m=1,2,3,\ldots,20}\{g_m(x)\}, \\ g_1(x) &= \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1n}x_n - 1 \leqslant 0, \\ g_2(x) &= \alpha_{21}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{2n}x_n - 1 \leqslant 0, \\ &\ldots \\ g_m(x) &= \alpha_{m1}x_1 + \alpha_{m2}x_2 + \ldots + \alpha_{mn}x_n - 1 \leqslant 0, \end{aligned} \tag{5.34}$$

где коэффициенты $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{mn}$ представлены матрицей (5.33). Результаты для данного примера представлены в Таблице 5.9. Как можно заметить, алгоритм 20 работает быстрее по сравнению с алгоритмом 17.

Некоторые примеры результатов экспериментов для задач с большими размерностями. В таблице 5.10 представлены результаты алгоритма 20 для $n = 3 \cdot 10^5$ размерностей. Из-за большой размерности оказалось невозможным получить результаты для алгоритма 17

Таблица 5.9. Сравнение результатов алгоритмов, пример 5.6.9.

ε	Итарации	Время, с	Итарации	Время, с
	Алгоритм 17		Алгоритм 20	
$1/2$	—	>300	16	0.158
$1/4$	—	>300	64	0.575
$1/6$	—	>300	144	1.089
$1/8$	—	>300	256	1.848

Таблица 5.10. Некоторые результаты алгоритма 20 для $n = 3 \cdot 10^5$.

ε	Итерации	Время, с	Итарации	Время, с	Итерации	Время, ММ:СС
	Пример 5.6.7		Пример 5.6.9		Пример 5.6.10	
$1/2$	16	34	16	30	16	35
$1/4$	64	123	64	118	64	141
$1/6$	144	278	144	272	144	326

и его модифицированной версии. Причина в том, что используемое в экспериментах программное обеспечение не может обработать входные данные из-за ошибки целочисленного переполнения. Оказывается, что выполнение алгоритма 20 не приводит к такой ошибке.

Пример 5.6.10. Пример геометрической задачи с квазивыпуклым целевым функционалом. Предположим, нам дано несколько точек A_k (центры шаров ω_k). Нужно найти шар наименьшего радиуса R охватывающего эти точки. Другими словами, необходимо найти центр такого шара, чтобы максимальное расстояние от центра до этих точек было наименьшим. В то же время мы предполагаем, что точка (центр) Q может лежать на некотором множестве, которое определяется функциональным ограничением (5.34), где коэффициенты $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{mn}$ представлены матрицей (5.33). Расстояние от Q до каждой из неподвижных точек A_k определяется следующим образом ($\rho > 1$):

$$d(Q, A_k) = \begin{cases} QA_k + (\rho - 1)r_k, & \text{если } |QA_k| > r_k \ (r_k \text{ — радиус } \omega_k), \\ \rho QA_k, & \text{в противном случае,} \end{cases}$$

где $d(Q, A_k) =: f(x)$ — вогнутая функция ($M_f = \rho$). Стоит отметить, что $d(Q, A_k)$ негладкая в таких точках Q, что $|QA_k| = r_k$. Для точек негладкости мы используем некоторый элемент субдифференциала Кларка

Таблица 5.11. Сравнение результатов алгоритмов, Пример 5.6.10.

ε	Итерации	Время, с	Итерации	Время, с
	Алгоритм 17		Алгоритм 20	
$1/2$	32680	199	16	0.095
$1/4$	65392	392	64	0.391
$1/6$	98135	587	144	0.862
$1/8$	—	>1000	256	1
$1/10$	—	>1000	400	2
$1/12$	—	>1000	576	3

в качестве аналога субградиента. Укажем и другие входные данные: $n = 1000$, $\rho = 2$, $x^0 = \frac{(0.1,\ldots,0.1)}{\|(0.1,\ldots,0.1)\|}$. Координаты точек A_k выбираются таким образом, что $\|A_k\| \in [1;2]$, количество точек A_k равно 1000 и $r_k = 1$ для всех $k = \overline{1,100}$. Результаты экспериментов для рассматриваемого примера представлены в таблице 5.11. Как можно заметить, алгоритм 20 работает быстрее, чем алгоритм 17. Однако скорость оценки в отношении целевой функции одинакова, но в отношении ограничений может быть намного хуже.

5.7 Алгоритмы зеркального спуска для задач выпуклой оптимизации с функциональными ограничениями: относительная липшицевость и относительная точность

В настоящем разделе мы рассмотрим применимость зеркальных спусков с переключениями к двум специальным классам задач. В первом подпункте мы покажем, как можно видоизменить алгоритм 19, чтобы возможно было получить оценки скорости сходимости для задач выпуклого программирования на классах относительно липшицевых целевого функционала и функционала ограничения (об этом была речь в начале настоящей главы). Второй подпункт посвящен двойственному методу для негладких задач с ограничениями. Далее будет показано, как возможно применять разработанные ранее адаптивные зеркальные

спуски с переключениями (алгоритмы 19 и 20) для задач минимизации выпуклых однородных функционалов с относительной точностью при наличии выпуклых функциональных ограничений. В зависимости от условия во втором подпункте настоящего раздела на субдифференциал однородного целевого функционала в нуле рассматривается два подхода, для каждого из которых получены оценки скорости сходимости. Если 0 есть внутренняя точка субдифференциала целевого функционала в нулевой точке, то к выделенному классу задач вполне возможно применять разработанные адаптивные методы при подходящем выборе входных параметров (ε и Θ_0). Для случая, когда 0 есть внутренняя точка субдифференциала целевого функционала в нуле, введено обобщение указанного условия с использованием аппарата полунормированных конусов. Для такого предположения также показано, как можно использовать зеркальные спуски с переключениями (с применением результатов, полученных на классе относительно липшицевых функционалов) и выписать оценки скорости сходимости. Отметим, что в случае $0 \notin Q$ (Q — допустимое множество задачи) для второго подхода важны результаты о метризуемости полунормированных конусов, а также условия разрешимости минимизационных задач в абстрактных нормированных и полунормированных конусах. Некоторые из таких результатов обсуждаются в последних трёх подпунктах данного раздела.

5.7.1 Относительная липшицевость: алгоритмы и оценки скорости сходимости В данном пункте мы рассмотрим некоторые методы для задачи минимизации выпуклой негладкой функции $f(x) \to \min$, но уже с относительно липшицевым целевым функционалом и функциональным ограничением $g(x) \leqslant 0$. Попрежнему, мы обозначаем через $\varepsilon > 0$ фиксированную точность, x_0 — начальное приближение такое, что для некоторого $\Theta_0 > 0$ верно неравенство $V(x_*, x^0) \leqslant \Theta_0^2$ (x_* — ближайшее к x^0 решение рассматриваемой задачи выпуклого программирования).

Начнём с того, что отметим, возможность предложить аналогичный алгоритму 19 метод и для задач со следующими релаксациями известного неравенства Коши-Буняковского для субградиентов $\nabla f(x)$ и $\nabla g(x)$:

$$\langle \nabla f(x), x - y \rangle \leqslant \omega \left\| \nabla f(x) \right\|_* \sqrt{2V(y,x)} \quad \forall y \in Q \tag{5.35}$$

и произвольного $x \in Q$: $V(x_*, x) \geqslant \frac{\varepsilon^2}{2}$, некоторой фиксированной

постоянной $\omega > 0$ (здесь x_* — ближайшее решение к начальной точке x_0 с точки зрения дивергенции V), а также

$$\langle \nabla g(x), x - y \rangle \leqslant M_g \sqrt{2V(y,x)} \tag{5.36}$$

для константы $M_g > 0$.

Ясно, что при $V(y,x) \geqslant \frac{1}{2}\|y-x\|^2$ неравенство (5.35) верно для $\omega = 1$. Неравенство (5.36), в частности, верно в случае относительной липшицевости g [125]. Теперь рассмотрим следующий алгоритм 27.

Алгоритм 27 Адаптивный зеркальный спуск, относительная липшицевость

Require: $\varepsilon > 0,\ \Theta_0 : d(x_*) \leqslant \Theta_0^2$

$x^0 = argmin_{x \in Q} d(x)$

$I =: \emptyset$

$k \leftarrow 0$

repeat

  if $g(x^k) \leqslant \varepsilon M_g$ **then**

    $h_k^f \leftarrow \frac{\varepsilon}{||\nabla f(x^k)||_*^2}$

    $x^{k+1} \leftarrow Mirr_{x^k}(h_k \nabla f(x^k))$ // *"продуктивные шаги"*

    $k \to I$

  else

    $h_k^g \leftarrow \frac{\varepsilon}{M_g}$

    $x^{k+1} \leftarrow Mirr_{x^k}(h_k \nabla g(x^k))$ // *"непродуктивные шаги"*

  end if

  $k \leftarrow k + 1$

until

$$\frac{2\Theta_0^2}{\varepsilon^2} \leqslant \sum_{k \in I} \frac{\omega^2}{||\nabla f(x^k)||_*^2} + |J|, \tag{5.37}$$

где $|J|$ — количество непродуктивных шагов (мы обозначим через $|I|$ количество продуктивных шагов, то есть $|I| + |J| = N$).

Ensure: $\widehat{x} = \frac{1}{\sum\limits_{k \in I} h_k^f} \sum\limits_{k \in I} h_k^f x^k$.

С учётом лемм 5.0.1 и 5.0.2 из (5.35) и (5.36) имеем:

$$h_k \left\langle \nabla f(x^k), x^k - x_* \right\rangle \leqslant \frac{\omega^2 \varepsilon^2}{2\left\|\nabla f(x^k)\right\|_*^2} + V(x_*, x^k) - V(x_*, x^{k+1}) \ \ \forall k \in I,$$

$$\varepsilon^2 < h_k g(x^k) \leqslant h_k \left\langle \nabla g(x^k), x^k - x_* \right\rangle \leqslant \frac{\varepsilon^2}{2} + V(x_*, x^k) - V(x_*, x^{k+1}) \ \ \forall k \in J.$$

После суммирования указанных неравенств получаем:

$$\sum_{k \in I} h_k \left(f(x^k) - f(x_*) \right) \leqslant \frac{\omega^2 \varepsilon}{2} \sum_{k \in I} h_k - \frac{\varepsilon^2}{2} |J| + V(x_*, x^0).$$

Поэтому, справедлива следующая

Теорема 5.7.1. *После выполнения критерия остановки* (5.37) *справедлива оценка:*

$$f(\widehat{x}) - f(x_*) \leqslant \omega^2 \varepsilon \ \textit{и} \ g(\widehat{x}) \leqslant \varepsilon M_g$$

или $V(x_*, x^k) < \frac{\varepsilon^2}{2}$ *для некоторого* k.

Замечание 5.7.2. Вместо условия (5.35) можно рассмотреть и неравенство

$$\langle \nabla f(x), x - y \rangle \leqslant M_f \sqrt{2V(y, x)},$$

которое верно в случае относительной липшицевости f [125]. Для этого необходимо выбирать в алгоритме 27 продуктивные шаги $h_k^f = \frac{\varepsilon}{M_f^2}$, а также критерий остановки

$$2V(x_*, x^0) = 2\Theta_0^2 \leqslant \frac{\varepsilon^2 |I|}{M_f^2} + \varepsilon^2 |J|.$$

После выполнения этого критерия будет верно неравенство

$$f(\widehat{x}) - f(x_*) \leqslant \varepsilon \text{ и } g(\widehat{x}) \leqslant \varepsilon M_g.$$

5.7.2 Минимизация выпуклой однородной функции с относительной точностью: двойственный метод для негладких задач с ограничениями

Будем рассматривать задачу

$$f(x) \to \min_{x \in Q} \tag{5.38}$$

минимизации выпуклой однородной функции [2)] на выпуклом замкнутом подмножестве $Q \subset \mathbb{R}^n$. В данном разделе работы будем рассматривать только евклидовы нормы $\|\cdot\|_2$ и по аналогии с ([38], глава 6)

[2)] Здесь имеется ввиду положительная однородность. Но мы далее условимся использовать термин «выпуклая однородная функция» вслед за главой 6 из диссертации Ю.Е. Нестерова [38].

предполагать

$$\mathrm{dom} f \equiv \mathbb{R}^n, \quad 0 \in \mathrm{int}\partial f(0),$$

откуда следует

$$\gamma_0 \|x\|_2 \leqslant f(x) \leqslant \gamma_1 \|x\|_2 \quad \forall\, x \in Q.$$

Далее в рассуждениях будет использована проекция $\bar{x}$ начала координат на множество Q относительно рассматриваемой (в данном разделе евклидовой) нормы

$$\|\bar{x}\|_2 := \min_{x \in Q}\{\|x\|_2\}.$$

Напомним, что в ([38], теорема 6.1.1) при сделанных предположениях обосновано следующее неравенство

$$\|\bar{x} - x_*\|_2 \leqslant \frac{2}{\gamma_0} f^*. \tag{5.39}$$

Будем рассматривать двойственную к (5.43) задачу

$$\min_{x \in Q}\{f(x) + \lambda g(x)\} = \min_{x \in Q}\{f(x) + \sum_{p=1}^{m} \lambda_p g_p(x)\} \to \max_{\lambda \in \mathbb{R}^m_+},$$

где $g(x)$ — вектор, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ — набор неотрицательных двойственных множителей. Сделаем предположение о том, что при всяком фиксированном наборе двойственных множителей λ существует и доступно значение $x(\lambda) \in arg\min_{x \in Q}\{f(x) + \lambda g(x)\}$. Например, это выполнено в случае возможности явно выписать двойственную задачу.

Для всякого вектора $a = (a_1, a_2, ..., a_n)$ условимся обозначать

$$[a]_+ = (\max\{0, a_1\}, \max\{0, a_2\}, ..., \max\{0, a_n\}).$$

Будем применять к двойственной задаче стандартный метод проекции субградиента:

$$\lambda^{k+1} = \left[\lambda^k + h_k \frac{g(x^k)}{\|g(x^k)\|_2}\right]_+, \tag{5.40}$$

где $x^k := x(\lambda^k)$ при $k \geqslant 0$, начальное приближение — $\lambda^0 = (0, ..., 0)$. Выберем в (5.40) постоянный шаг

$$h_k = \frac{R}{\sqrt{N}} \quad (k = 0, 1, \ldots, N), \tag{5.41}$$

где $R \leqslant \|\bar{x} - x_*\|_2$. Заметим, что если на каком-то шаге $g(x^k) = 0$, то это означает, что двойственная задача уже точно решена на k-й итерации и процесс можно остановить. Поэтому, не уменьшая общности рассуждений, случай $g(x^k) = 0$ можно исключить из рассмотрения. Справедлива следующая

Теорема 5.7.3. *Пусть для некоторой постоянной $L_g > 0$ верно неравенство $\|g(x)\|_2 \leqslant L_g$ при любом $x \in Q$ и количество итераций N метода* (5.40) *с постоянным шагом* (5.41) *удовлетворяет неравенству*

$$N \geqslant \frac{4L_g^2}{\gamma_0^2 \delta^2}. \tag{5.42}$$

Тогда после выполнения N итераций метода (5.40) *заведомо будут верны неравенства:*

$$f(\hat{x}) \leqslant f^*(1+\delta), \qquad \|[g(\hat{x})]_+\|_2 \leqslant \frac{L_g}{\sqrt{N}},$$

где

$$\hat{x} = \sum_{k=0}^{N-1} \mu_k x^k, \quad \mu_k = \frac{h_{N-1}\|g(x^{N-1})\|_2^{-1}}{\sum_{j=0}^{k} \frac{h_j}{\|g(x^j)\|_2}} = \frac{\|g(x^{N-1})\|_2^{-1}}{\sum_{j=0}^{k} \frac{1}{\|g(x^j)\|_2}},$$

$$k = 0, 1, 2, ..., N-1.$$

Доказательство. Для метода (5.40) при условии $\|g(x)\|_2 \leqslant L_g$ справедлива оценка ([70], лемма 8.45):

$$f(\hat{x}) - f^* + R\|[g(\hat{x})]_+\|_2 \leqslant \frac{L_g}{2} \cdot \frac{R^2 + \sum\limits_{k=0}^{N-1} h_k^2}{\sum\limits_{k=0}^{N-1} h_k}$$

для некоторого числа $R > 0$ и начального приближения $\lambda^0 = (0, 0, \ldots, 0)$.

Тогда ввиду (5.45) справедливы неравенства:

$$\frac{R^2 + \sum\limits_{k=0}^{N-1} h_k^2}{\sum\limits_{k=0}^{N-1} h_k} = \frac{2R}{\sqrt{N}} \leqslant \frac{2\|\bar{x} - x_*\|_2}{\sqrt{N}} \leqslant \frac{4f^*}{\gamma_0\sqrt{N}}.$$

Поэтому $f(\hat{x}) - f^* + R\|[g(\hat{x})]_+\|_2 \leqslant \dfrac{2L_g f^*}{\gamma_0\sqrt{N}} \leqslant \delta f^*$ (напомним, что f^* изначально предполагается положительным), если N удовлетворяет (5.42). Поэтому после (5.42) итераций метода (5.40) гарантированно будет выполнено неравенство:

$$f(\hat{x}) \leqslant f^*(1+\delta),$$

причем $R\|[g(\hat{x})]_+\|_2 \leqslant \dfrac{RL_g}{\sqrt{N}}$ или $\|[g(\hat{x})]_+\|_2 \leqslant \dfrac{L_g}{\sqrt{N}}$. □

Замечание 5.7.4. Если допустимое множество задачи есть аффинное подпространство и $\bar{x}$ — проекция 0 на это подпространство, то согласно ([38], теорема 6.1.1) неравенство (5.45) можно усилить

$$\|\bar{x} - x_*\|_2 \leqslant \frac{1}{\gamma_0} f^*,$$

что позволит уменьшить оценку количества итераций (5.42) в 4 раза.

Очевидно, что аналогичный вывод можно сделать и в случае $0 \in Q$, если ограничения записывать в виде системы неравенств с выпуклыми функционалами, а точку $\bar{x}$ выбрать нулевой. В частности, для задачи (5.54) вполне возможно ограничение вида $x \in \mathcal{L}$ ($\mathcal{L}$ — аффинное подпространство) записать в виде системы двух неравенство для линейных функционалов ограничения, а в качестве Q выбрать всё пространство. Такой подход даст возможность по сравнению с [38], в частности, избежать при выборе начальной точки $\bar{x}$ дополнительной операции проектирования 0 на аффинное подпространство $\mathcal{L}$.

5.7.3 Оценки скорости сходимости зеркальных спусков для задач с относительной точностью

Оказывается, возможно применить алгоритм 19 для (вообще говоря, негладких) задач выпуклой однородной минимизации с относительной точностью. Такая постановка восходит к работам Ю. Е. Нестерова (см. главу 6 диссертации [38]). Как показано Ю. Е. Нестеровым, подход к оценке качества решения задачи с точки зрения именно относительной точности вполне оправдан для разных прикладных задач (линейное программирование, проектирование механических конструкций, задача оптимальной эллипсоидальной аппроксимации выпуклого множества и пр.) при подходящих условиях на желаемую относительную точность. Известно, что достаточно широкий класс задач оптимизации с относительной

точностью можно сводить к минимизации выпуклой однородной функции. Итак, рассматривается на выпуклом замкнутом множестве $Q \subset \mathbb{R}^n$ задача минимизации выпуклой однородной функции вида

$$f(x) \to \min_{x \in Q} \tag{5.43}$$

с выпуклыми функционалами ограничений

$$g_p(x) \leqslant 0, \quad p = \overline{1, m}.$$

Стандартно будем обозначать $g(x) := \max_{1 \leqslant p \leqslant m}\{g_p(x)\}$.

Рассмотрим сначала предположение о том, что 0 есть внутренняя точка субдифференциала $\partial f(0)$:

$$0 \in \operatorname{int} \partial f(0). \tag{5.44}$$

Как известно (см. теорему 6.1.1 из [38]), в таком случае для некоторого $\gamma_0 > 0$ верно неравенство

$$\|x^0 - x_*\| \leqslant \frac{2}{\gamma_0} f^*. \tag{5.45}$$

Теорема 5.7.5. *Пусть выпуклый однородный функционал f M_f-липшицев на Q при некотором $M_f > 0$ и удовлетворяет условию (5.44), g — выпуклый функционал на Q, а также для некоторого $C > 0$ верно*

$$2V(x_*, x^0) \leqslant C^2 \|x_* - x^0\|^2 \tag{5.46}$$

и начальная точка x^0 выбрана так, что $\|x^0\| := \min_{x \in Q}\{\|x\|\}$. Тогда для всякого $\delta > 0$ можно подобрать входные параметры $\varepsilon > 0$ и $\Theta_0 > 0$ алгоритма 19 так, что после

$$N \geqslant \frac{4C^2 \max\{1, M_f^2\}}{\gamma_0^2 \delta^2} \tag{5.47}$$

итераций этого метода гарантированно будут выполнены следующие неравенства:

$$f(\widehat{x}) \leqslant f^*(1 + \delta), \qquad g(\widehat{x}) \leqslant \delta f^* \|\nabla g(\widehat{x})\|_*.$$

Доказательство. В силу предположения (5.46) возможно подобрать входной параметр Θ_0 алгоритма 19 так, чтобы для некоторого $R > 0$ было верно

$$2V(x_*, x^0) \leqslant 2\Theta_0^2 = R^2 \leqslant C^2 \|x_* - x^0\|^2.$$

Выберем теперь $\varepsilon = \dfrac{R\gamma_0\delta}{2C}$. Тогда после остановки алгоритма 19 согласно теореме 5.2.1 будут выполняться следующие неравенства:

$$f(\widehat{x}) - f^* \leqslant \frac{R\gamma_0\delta}{2C}, \qquad g(\widehat{x}) \leqslant \frac{R\gamma_0\delta\|\nabla g(\widehat{x})\|_*}{2C}.$$

Теперь с учетом сделанного предположения $R \leqslant C\|x^0 - x_*\|$ и в силу неравенства (5.45) имеем

$$f(\widehat{x}) - f^* \leqslant \frac{R\gamma_0\delta}{2C} \leqslant \frac{\gamma_0\delta\|x_* - x^0\|}{2} \leqslant f^*\delta,$$

а также $g(\widehat{x}) \leqslant \delta f^*\|\nabla g(\widehat{x})\|_*$.

Вспоминая оценку количества итераций (5.18) (после которых гарантированно будет выполнен критерий остановки алгоритма 19) с учетом $\varepsilon = \dfrac{R\gamma_0\delta}{2C}$, получим оценку (5.47) числа итераций алгоритма 19), после которых для точки-выхода заведомо будет достигнута δ-относительная точность решения поставленной задачи по целевому функционалу. $\square$

Замечание 5.7.6. Отметим [36, 138], что предположение (5.46) вполне естественно для многих известных примеров выбора прокс-функций вне зависимости от x^0 и x_*. Хорошо известно [36, 138], что если для произвольных $x, y \in Q$ существует зависящая от величины размерности пространства константа $C_n > 0$ (причём $C_n = O(\log n)$) такая, что $d(x - y) \leqslant C_n\|x - y\|^2$, то $V(x, y) \leqslant C_n\|x - y\|^2$ при любых $x, y \in Q$. Тогда (5.46) будет верно при всяком $C^2 \geqslant \frac{C_n}{2}$ вне зависимости от x^0 и x_*.

Замечание 5.7.7. Если воспользоваться адаптивным критерием остановки алгоритма 19 (пункт 14 листинга) и не заменять $\displaystyle\sum_{k\in I} \frac{1}{\|\nabla f(x^k)\|^2}$ на $\dfrac{|I|}{M_f^2}$, то при практической реализации метода достижение приемлемого качества решения возможно за меньшее число итераций по сравнению с доказанной оценкой (5.47). В частности, неравенства (5.18) и (5.47) возможно заменить на следующие:

$$N \geqslant \frac{2\Theta_0^2}{\varepsilon^2}\max\left\{1, \max_{k\in I}\|\nabla f(x^k)\|_*^2\right\} = \frac{4C^2}{\gamma_0^2\delta^2}\max\left\{1, \max_{k\in I}\|\nabla f(x^k)\|_*^2\right\}.$$

Эти оценки могут оказаться лучшими по сравнению с (5.18) и (5.47) на классе M_f-липшицевых целевых функционалов.

Замечание 5.7.8. Рассмотрим теперь случай евклидовой нормы и прокс-структуры, т.е.

$$V(x,y) = d(x-y) = \frac{1}{2}\|x-y\|_2^2$$

для произвольных $x, y \in Q$ и предположим, что допустимое множество задачи есть аффинное подпространство, начальная точка x^0 есть евклидова проекция начала координат на допустимое множество задачи. В этом случае рассуждения существенно опираются на неравенство

$$\|x_* - x^0\|_2 \leqslant \frac{1}{\gamma_0} f^*,$$

для выпуклого и однородного функционала f, что позволяет уменьшить константу в полученной оценке (5.47). Кроме того, в этом случае нет необходимости вводить предположение (5.46). Действительно, для евклидовой нормы возможно подобрать входной параметр Θ_0 алгоритма 19 так, чтобы для некоторых $R > 0$ и $C \geqslant 1$ было верно

$$\|x^0 - x_*\|_2^2 \leqslant 2\Theta_0^2 = R^2 \leqslant C^2 \|x^0 - x_*\|^2$$

и положить $\varepsilon = \dfrac{R\gamma_0\delta}{C}$. Это означает, что после остановки алгоритма 19 будут выполняться неравенства

$$f(\widehat{x}) - f^* \leqslant \frac{R\gamma_0\delta}{C}, \qquad g(\widehat{x}) \leqslant \frac{R\gamma_0\delta\|g(\widehat{x})\|_*}{C}.$$

Поэтому для описанных выше предположений имеем

$$f(\widehat{x}) \leqslant f^*(1+\delta), \qquad g(\widehat{x}) \leqslant \delta f^* \|g(\widehat{x})\|_*.$$

При этом в силу выбора $\varepsilon = \dfrac{R\gamma_0\delta}{C}$ оценка числа итераций (5.47), необходимых для выполнения критерия остановки алгоритма, уменьшается в 4 раза. В отличие от предыдущего результата параметр $C \geqslant 1$ можно выбрать произвольно. Чем больше выбирается параметр C, тем выше будет точность достигнутого решения, но вычислительные затраты (необходимое количество итераций) будут большими. Ясно, что улучшение оценки (5.47) в 4 раза в условиях теоремы 5.7.5 возможно и для произвольной нормы, если $0 \in Q$.

Далее, покажем, как можно выписать оценки сходимости для зеркальных спусков с переключениями в случае, когда 0 не обязательно

есть внутренняя точка субдифференциала f в нуле. Введём следующий ослабленный вариант этого условия

$$B_{\gamma_0}^{K^*}(0) \subseteq \partial f(0) \subseteq B_{\gamma_1}^{K^*}(0), \tag{5.48}$$

где K^* — сопряженный конус к некоторому полунормированному конусу $K \subset \mathbb{R}^n$ с законом сокращения и конус-полунормой $\|\cdot\|_K$ (отличие от обычной полунормы в том, что $\|\alpha x\|_K = \alpha\|x\|_K$ лишь для $\alpha \geqslant 0$). Здесь под *сопряженным конусом* K^* понимается набор функционалов вида $\psi_\ell = \max\{0, \ell(x)\}$ для линейных функционалов $\ell : K \to \mathbb{R} : \ell(x) \leqslant C_\ell\|x\|_K$ при некотором $C_\ell > 0\ \forall x \in K$. Ясно, что K^* будет выпуклым конусом с операциями сложения

$$\psi_{\ell_1} \oplus \psi_{\ell_2} := \psi : \quad \psi(x) = \max\{0, \ell_1(x) + \ell_2(x)\}$$

и умножения на скаляр $\lambda \geqslant 0$ $\psi_{\lambda\ell}(x) = \lambda\psi_\ell(x) = \lambda\max\{0, \ell(x)\} \quad \forall x \in K$. На K^* можно ввести норму $\|\psi_\ell\|_{K^*} = \sup_{\|x\|_K \leqslant 1}\max\{0, \ell(x)\} = \sup_{\|x\|_K \leqslant 1}\ell(x)$ и шар радиуса r $B_r^{K^*}(0) = \{\psi_\ell \in K^* \mid \|\psi_\ell\|_{K^*} \leqslant r\}$.

Из аналога теоремы об опорном функционале в нормированных конусах (см. следствие 5.7.18 из теоремы 5.7.16 далее) получаем, что

$$\|x\|_K = \max_{\psi_\ell \in B_1^{K^*}(0)} \ell(x). \tag{5.49}$$

Для полунормированного конуса при $\|x\|_K = 0$ достаточно выбрать $\ell \equiv 0$ и (5.49) также будет верно. Приведем некоторый пример пары (K, K^*).

Пример 5.7.9. Пусть $K = \{(x, y) \mid x, y \in \mathbb{R}\}$ и $\|(x, y)\|_K = \sqrt{x^2 + y^2} + y$. Можно проверить, что в таком случае

$$K^* = \{\psi_\ell \mid \ell((x, y)) = \lambda x + \mu y : \mu + \frac{\lambda^2}{\mu} < +\infty \text{ или } \lambda = \mu = 0\}, \text{ а}$$

$$\|\psi_\ell\|_{K^*} = \begin{cases} 0, & \text{если } \lambda = \mu = 0; \\ \dfrac{\mu}{2} + \dfrac{\lambda^2}{2\mu}, & \text{если } \mu + \dfrac{\lambda^2}{\mu} < +\infty \text{ при } \mu > 0. \end{cases}$$

Тогда $B_1^{K^*}(0)$ имеет вид круга на плоскости (λ, μ) радиуса 1 с центром в точке $\lambda = 0$, $\mu = 1$. Не уменьшая общности рассуждений, будем полагать $K = \bigcup\limits_{r \geqslant 0} B_r^K(0)$, а также $x_* \in K$ для точного решения x_* рассматриваемой задачи минимизации f на Q.

Согласно схеме рассуждений ([38], глава 6) для вывода оценок скорости сходимости методов с относительной точностью необходимо знать оценку R величины расстояния от точки старта x^0 до ближайшего решения x_*. Однако в конусах, вообще говоря, не задана операция вычитания и поэтому в качестве аналога нормы разности можно использовать метрику $d^K(x^0, x_*)$, где

$$d^K(x, y) = \sup_{\|\psi_\ell\|_{K^*} \leqslant 1} |\psi_\ell(x) - \psi_\ell(y)| \quad \forall x, y \in Q.$$

Сделаем предположение о существовании такой метрики. Некоторые условия, при которых нормированный конус допускает существование метрики такого типа, приведены в последующих пунктах данного раздела работы.

Можно получить следующий аналог теоремы 6.1.1 [38] для указанного выше предположения ($x^0, x_* \in K$). Пусть x^0 таково, что

$$\|x^0\|_K = \min\{\|x\|_K \,|\, x \in Q\}. \tag{5.50}$$

Теорема 5.7.10. *Пусть верно* (5.48). *Тогда*

1) $\forall\, x \in K \quad \gamma_0\|x\|_K \leqslant f(x) \leqslant \gamma_1\|x\|_K$. *Более того,*

$$\frac{\gamma_0}{\gamma_1} f(x^0) \leqslant \gamma_0\|x^0\|_K \leqslant f(x_*) \leqslant f(x^0) \leqslant \gamma_1\|x^0\|_K.$$

2) *Для всякого точного решения $x_* \in K$ справедливо неравенство:*

$$d^K(x^0, x_*) \leqslant \|x^0\|_K + \|x_*\|_K \leqslant \frac{2}{\gamma_0} f^* \leqslant \frac{2}{\gamma_0} f(x^0). \tag{5.51}$$

Доказательство. В силу выпуклости и однородности f имеем:

$$f(x) = \max_{\upsilon}\{\langle \upsilon, x\rangle \,|\, \upsilon \in \partial f(0)\},$$

откуда $\max\limits_{\upsilon}\{\langle \upsilon, x\rangle \,|\, \upsilon \in B^{K^*}_{\gamma_0}(0)\} \leqslant f(x) \leqslant \max\limits_{\upsilon}\{\langle \upsilon, x\rangle \,|\, \upsilon \in B^{K^*}_{\gamma_1}(0)\}$. Согласно (5.49) последнее неравенство означает, что $\gamma_0\|x\|_K \leqslant f(x) \leqslant \gamma_1\|x\|_K \,\forall\, x \in K$. Далее,

$$f^* = \min_x\{f(x) \,|\, x \in Q\} \geqslant \min_x\{\gamma_0\|x\|_K \,|\, x \in Q\} = \gamma_0\|x^0\|_K,$$

т.е. $\alpha f(x^0) \leqslant \gamma_0\|x^0\|_K \leqslant f^*$.

Поскольку $x_* \in K$, то $f^* = f(x_*) \geqslant \gamma_0 \|x_*\|_K$, то

$$d^K(x_0, x_*) \leqslant \sup_{\|\psi_\ell\|_* \leqslant 1} \max\{0, \ell(x^0)\} + \sup_{\|\psi_\ell\|_* \leqslant 1} \max\{0, \ell(x_*)\} =$$

$$= \|x^0\|_K + \|x_*\|_K \leqslant \frac{2}{\gamma_0} f^*,$$

что и требовалось. □

Замечание 5.7.11. Если $0 \in Q$, то в качестве начальной точки можно просто выбрать $x^0 = 0$, и тогда неравенство (5.51) принимает упрощенный вид

$$d^K(0, x_*) = \|x_*\|_K \leqslant \frac{1}{\gamma_0} f^* \leqslant \frac{1}{\gamma_0} f(x^0). \tag{5.52}$$

Замечание 5.7.12. Условие $x_* \in K$ может оказаться не существенным. Например, в рассмотренном выше примере 5.7.9 K совпадает со всем пространством.

Обсудим теперь применимость алгоритма 19 к поставленной задаче выпуклого программирования с относительной точностью по целевому функционалу. Для этого достаточно выбрать прокс-структуру так, чтобы

$$V(x_*, x^0) \leqslant \widehat{\omega} d^K(x_*, x^0),$$

для некоторой постоянной $\widehat{\omega} > 0$. Ясно, что для некоторой постоянной $\widehat{\omega} > 0$ этого всегда можно добиться.

Поскольку функционал f однороден, то при условии ограниченности субдифференциала $\partial f(0)$ для некоторой константы $M_f > 0$ будет верно $\|\nabla f(x)\|_* \leqslant M_f$. Поэтому критерий остановки алгоритма 19 ввиду (5.18) заведомо будет выполнен после $2\Theta_0^2 \max\left\{1, M_f^2\right\} \varepsilon^{-2}$ итераций. Будем полагать, что при некотором $\widehat{\omega} > 0$

$$\Theta_0^2 = \widehat{\omega} d^K(x^0, x_*) \geqslant V(x_*, x^0). \tag{5.53}$$

Далее, полагаем, что вместо метрики d^K используется $\widehat{\omega} d^K$ с соответствующим подбором $\gamma_0 > 0$. Для некоторого параметра $\widehat{N}$ выберем $\varepsilon = \frac{\Theta_0^2}{\sqrt{\widehat{N}}}$. При данном ε после алгоритма 19 будут верны неравенства $f(\widehat{x}) - f(x_*) \leqslant \frac{\Theta_0^2}{\sqrt{\widehat{N}}}$ и $g(\widehat{x}) \leqslant \frac{M_g \Theta_0^2}{\sqrt{\widehat{N}}}$. Теперь, по теореме 5.7.10 имеем:

$$\frac{\Theta_0^2}{\sqrt{\widehat{N}}} \leqslant \frac{2 f(x_*)}{\gamma_0 \sqrt{\widehat{N}}}, \text{ то есть } f(\widehat{x}) \leqslant f(x_*) \left(1 + \frac{2}{\gamma_0 \sqrt{\widehat{N}}}\right) \text{ и } g(\widehat{x}) \leqslant \frac{M_g \Theta_0^2}{\sqrt{\widehat{N}}}.$$

Поэтому для достижения относительной точности $\delta > 0$ по функции заведомо достаточно $\widehat{N} \geqslant \frac{4}{\gamma_0^2\delta^2}$. Отметим, что критерий остановки алгоритма 19 при этом заведомо выполнен после $\left\lceil \frac{8\max\left\{1, M_f^2\right\}}{\gamma_0^2\delta^2\Theta_0^2} \right\rceil$ итераций алгоритма 19, поскольку

$$\frac{2\Theta_0^2\max\left\{1, M_f^2\right\}}{\frac{\Theta_0^4}{\widehat{N}}} = \frac{2\widehat{N}\max\left\{1, M_f^2\right\}}{\Theta_0^2} \leqslant \left\lceil \frac{8\max\left\{1, M_f^2\right\}}{\gamma_0^2\delta^2\Theta_0^2} \right\rceil,$$

если $\widehat{N} = \frac{4}{\gamma_0^2\delta^2}$. Поэтому верна следующая

Теорема 5.7.13. *Пусть однородный выпуклый функционал f M_f-липшицев на Q для некоторой $M_f > 0$. Тогда после $\left\lceil \frac{8\max\left\{1, M_f^2\right\}}{\gamma_0^2\delta^2\Theta_0^2} \right\rceil$ итераций модифицированного алгоритма 19 гарантированно будут выполнены неравенства:*

$$f(\widehat{x}) \leqslant f(x_*)(1+\delta) \text{ и } g(\widehat{x}) \leqslant \frac{M_g\Theta_0^2\gamma_0\delta}{2}.$$

Ясно, что можно выбрать достаточно большое $\Theta_0^2 \geqslant M_f^2$, и тогда оценка числа итераций упростится до $\frac{8}{\gamma_0^2\delta^2}$. Если $\Theta_0^2 \geqslant 1$ (что вполне естественно, так как иначе x^0 и так близко к x_*), то оценку количества итераций возможно заменить на $\left\lceil \frac{8\max\left\{1, M_f^2\right\}}{\gamma_0^2\delta^2} \right\rceil$. Также вполне можно гарантировать достижения аналогичных вычислительных гарантий после выполнения критерия остановки алгоритма 19 для выбранного параметра ε. Возможно использование алгоритма 19 с постоянными шагами, что даст возможность использовать не обязательно сильно выпуклую прокс-функцию (см. раздел 5.7.1 выше).

Замечание 5.7.14. Отметим, что использование метрики d^K для оценки качества решения здесь связано с отсутствием требования, чтобы 0 была внутренней точкой $\partial f(0)$. Если 0 — внутренняя точка $\partial f(0)$, то вместо метрики d^K вполне возможно использовать обычное евклидово расстояние.

Также заметим, что даже если 0 не есть внутренняя точка $\partial f(0)$ и необходимо использовать метрику d^K для подходящего K, то для проверки иных необходимых условий на f (липшицевость)вполне можно использовать обычную евклидову прокс-структуру, подобрав при этом подходящий параметр $\widehat{\omega} > 0$ в (5.53).

Отметим, что схема рассуждений, которая привела нас к теореме 5.7.13, может быть применена (с некоторыми изменениями) для алгоритма 20. Пусть $\varepsilon = \frac{\Theta_0^2}{\sqrt{\widehat{N}}}$ для некоторого $\widehat{N}$, и тогда после $\left\lceil \frac{2\Theta_0^2}{\varepsilon^2} \right\rceil = \left\lceil \frac{2\widehat{N}}{\Theta_0^2} \right\rceil$ итераций алгоритма 20 будут верны неравенства $f(\widehat{x}) - f^* \leqslant \frac{M_f \Theta_0^2}{\sqrt{\widehat{N}}}$ и $g(\widehat{x}) \leqslant \frac{M_g \Theta_0^2}{\sqrt{\widehat{N}}}$.

По теореме 5.7.10 получаем $f(\widehat{x}) \leqslant f^* \left(1 + \frac{2M_f}{\gamma_0 \sqrt{\widehat{N}}}\right)$. Это означает, что для достижения δ-относительной точности решения будет достаточно $\widehat{N} \geqslant \frac{4M_f^2}{\gamma_0^2 \delta^2}$ или $\left\lceil \frac{8M_f^2}{\gamma_0^2 \delta^2 \Theta_0^2} \right\rceil$ итераций алгоритма 20. При этом с точки зрения ограничения для выхода $\widehat{x}$ будет верно $g(\widehat{x}) \leqslant \frac{M_g \Theta_0^2 \delta \gamma_0}{2M_f}$.

Если $0 \in Q$, то ввиду неравенства (5.52) оценки необходимого числа итераций можно уменьшить в четыре раза для обоих рассматриваемых методов.

Отметим в завершении данного пункта, что предложенные схемы и полученные в настоящем пункте результаты для условных задач с относительной точностью вполне возможно применить и к рассмотренной в ([38], глава 6) постановки задачи нахождения

$$f^* = \min_x \{f(x) : x \in \mathcal{L}\}, \quad \mathcal{L} = \{x \in \mathbb{R}^n : Cx = b\}, \tag{5.54}$$

где выпуклая функция $f(x)$ однородная степени 1, C — $p \times n$-матрица (не ограничивая общности можно считать, что матрица C имеет полный строчный ранг) и вектор b отличен от 0. В таком случае для задачи (5.54) вполне возможно ограничение вида $x \in \mathcal{L}$ ($\mathcal{L}$ — аффинное подпространство) записать в виде системы двух неравенств с линейными функционалами ограничения, а в качестве Q выбрать всё пространство. Такой подход даст возможность по сравнению с [38], в частности, избежать при выборе начальной точки операции проектирования 0 на аффинное подпространство $\mathcal{L}$.

5.7.4 О некоторых классах метризуемых выпуклых конусов с нормами

В завершение данного раздела приведём некоторые использованные в предыдущем пункте сведения о выпуклых нормированных конусах. Особое выделим условия существования на таких конусах метрики. Метризуемость используемого полунормированного конуса существенно важна при обосновании теоремы 5.7.13

выше. Возможна ситуация, когда нет возможности записать ограничения так, что $0 \in Q$ и для выбора начальной точки x^0 необходимо решать вспомогательную подзадачу (5.50). Тогда важны также условия разрешимости таких задач в нормированных и полунормированных конусах, некоторые из которых мы также опишем ниже.

Выпуклые нормированные и полунормированные конусы. Во многих задачах возникают так называемые выпуклые конусы — структуры, похожие на линейные пространства, но с умножением лишь на неотрицательные скаляры $\lambda \geqslant 0$. *Абстрактными выпуклыми конусами* или *выпуклыми конусами* принято называть набор элементов Q с заданными операциями сложения, а также умножения на неотрицательный скаляр, причём Q — коммутативная полугруппа по сложению и для произвольных чисел $\lambda, \mu \geqslant 0$, а также элементов $x, y \in Q$ верны соотношения:

$$1 \cdot x = x; \quad (\lambda\mu)x = \lambda(\mu x); \quad 0 \cdot x = 0;$$

$$\lambda(x + y) = \lambda x + \lambda y; \quad (\lambda + \mu)x = \lambda x + \mu x.$$

Попутно, как правило, требуется выполнение так называемого *закона сокращения* («cancellation law»):

$$x + y = y + z \Longleftrightarrow x = z \quad \forall x, y, z \in Q. \tag{5.55}$$

Выпуклыми конусами будут, в частности, наборы векторов с неотрицательными координатами, наборы неотрицательных функций, неубывающих функций с естественными операциями сложения и умножения на скаляр, а также наборы выпуклых компактных подмножеств банахова пространства со сложением по Минковскому. В некоторых выпуклых конусах Q возможно ввести *норму* $||\cdot|| : Q \to \mathbb{R}$:

$$||x|| \geqslant 0, \ ||x|| = 0 \Leftrightarrow x = 0; \ ||\lambda x|| = \lambda||x||; \ ||x + y|| \leqslant ||x|| + ||y|| \tag{5.56}$$

для всяких $x, y \in Q$ и произвольного $\lambda \geqslant 0$.

Выпуклые конусы с законом сокращения (5.55) и нормой, удовлетворяющей (5.56), мы будем называть *нормированными конусами*. Такие структуры активно исследовались многими авторами [43, 116, 161, 163, 165]. Отметим в частности работы [100, 116, 162, 163, 165], посвящённые теории двойственности в нормированных конусах с использованием неотрицательных линейных функционалов.

Нормированным конусом является всякое линейное пространство с несимметричной нормой $||\cdot|$. Несимметричная норма отличается от обычной тем, что вместо однородности требуется лишь положительная однородность $||\lambda x| = \lambda||x|$ при $\lambda \geqslant 0$. При этом, вообще говоря, $||-x| \neq ||x|$. Важный и естественный пример несимметричной нормы — функционал Минковского выпуклого несимметричного множества, содержащего 0. Несимметричные нормы в бесконечномерном анализе были введены М. Г. Кейном (см., например, [33]) в связи с известной проблемой моментов. Пространства с несимметричными нормами активно исследовались в последние десятилетия Е. П. Долженко, А. Р. Алимовым, П. А. Бородиным и другими отечественными математиками (см. [2,12,25] и имеющиеся там ссылки). Отметим также недавние работы в области несимметрично нормированных (и полунормированных) пространств Г. Е. Иванова и М. С. Лопушански [27, 108], а также монографию С. Кобзаша [86]. Известны исследования аппарата несимметричных норм в связи с задачами теоретической информатики Л. М. Гарсиа-Раффи, С. Ромагуера, Э. А. Санчез-Перез, О. Валеро и др. [100, 162].

В работах П. А. Бородина [12] и А. Р. Алимова [2] были получены аналоги теоремы Банаха–Мазура в классе пространств E с несимметричной нормой $\|\cdot|$, которые сепарабельны относительно согласованной с $\|\cdot|$ обычной симметричной нормы $\|x\|_{sym} = \max\{\|x|, \|-x|\}$. В частности, в работе [12] доказано линейное инъективное изометричное вложение всякого $\|\cdot\|_{sym}$-сепарабельного несимметрично нормированного пространства в пространство $C[0;1]$ непрерывных функций $f : [0;1] \to \mathbb{R}$ с несимметричной полунормой

$$\|f\| = \max_{0 \leqslant t \leqslant 1} \max\{0, f(t)\}.$$

В статье [2] был выделен более узкий класс *метризуемых* пространств с несимметричной нормой $\|\cdot|$, в которых эквивалентны нормы $\|\cdot|$ и $\|\cdot\|_{sym}$. В [2] доказано, что пространство с несимметричной нормой $(E, \|\cdot|)$ можно изометрически изоморфно вложить в $C([0,1])$ как аффинное линейное многообразие L тогда и только тогда, когда Q метризуемо и сепарабельно. Ввиду метризуемости E нормы $\|\cdot|$ и $\|\cdot\|_{sym}$ эквивалентны и поэтому безразлично, по какой из них рассматривать сепарабельность. Изометричность в [2] понимается в следующем смысле ($B_{C([0,1])}$ — единичный шар пространства $C([0,1])$): существует

точка $\theta \in L \cap int\, B_{C([0,1])}$ такая, что функционал Минковского множества $B_{C([0,1])} \cap L$ относительно θ совпадает с несимметричной нормой $||\cdot|$. По сути, в [2] в отличие от [12] доказано существование изометричного вложения несимметрично нормированного пространства в $C[0;1]$, но при этом вместо линейного вложения рассматривается аффинное. Также в [2] сужается класс рассматриваемых пространств: например, неметризуемо пространство непрерывных функций $f : [0;1] \to \mathbb{R}$ со следующей несимметричной нормой (см. [2]):

$$||f|| = \max_{0 \leqslant t \leqslant 1} \max\{0, f(t)\} + \int_0^1 |\min\{0, f(t)\}| dt.$$

В обзорном порядке приведём некоторые необходимые нам в дальнейшем понятия и результаты, среди которых полученный нами в [171] аналог теоремы Хана–Банаха о продолжении линейного функционала с подконуса на весь конус с сохранением выпуклой оценки. Также мы докажем аналог леммы об опорном функционале в нормированном конусе, в котором в отличие от ([171], следствие 3.1) необязательно $\|x\| = \|y\|$ при $x + y = 0$. Начнём с некоторых вспомогательных понятий. Сначала напомним, что неотрицательный функционал $p : Q \to \mathbb{R}$ называется *выпуклым*, если для любых $x, y \in Q$ и $\lambda \geqslant 0$ верно: $p(\lambda x) = \lambda p(x)$ и $p(x + y) \leqslant p(x) + p(y)$. В [171] мы ввели следующий аналог понятия подпространства в выпуклых конусах.

Определение 5.7.15. Будем говорить, что выпуклый конус Y есть *подконус* Q, если $Y \subset Q$, а также для всякого $x \in Q$ такого, что $x+y, y \in Y$ верно $x \in Y$.

Например, для фиксированного набора элементов $x_1, x_2, ..., x_n \in Q$ множество

$$Y = \left\{ x \in Q \,\middle|\, x + \sum_{k=1}^{n} \mu_k x_k = \sum_{k=1}^{n} \lambda_k x_k \text{ для всех } \lambda_k, \mu_k \geqslant 0,\ k = \overline{1,n} \right\}$$

будет подконусом Q.

Приведём теперь полученный нами в [171] аналог теоремы Хана–Банаха о продолжении линейного функционала с подконуса $Y \subset Q$ на весь конус Q в классе выпуклых конусов Q с законом сокращения.

Теорема 5.7.16. *Пусть Q — выпуклый конус с законом сокращения и $p : Q \to \mathbb{R}$ — выпуклый функционал на Q, Y — подконус Q, а также существует линейный функционал $\ell : Y \to \mathbb{R}$ с оценкой $\ell(y) \leqslant p(y)$ для любого $y \in Y$. Тогда существует такой линейный функционал $L : Q \to \mathbb{R}$, что $L(x) \leqslant p(x)$ для любого $x \in Q$ и $L(y) = \ell(y)$ при всех $y \in Y$.*

Доказательство. 1) Предположим, что для фиксированного элемента $e \in Q$

$$Q = Lin(e, Y) = \{x \in Q \mid x + \lambda_1 e + y_1 = \lambda_2 e + y_2$$

$$\text{для некоторых } y_1, y_2 \in Y \text{ и } \lambda_1, \lambda_2 \geqslant 0\}.$$

В силу $Q \in (CL)$ $x + \lambda_1 e + y_1 = \lambda_2 e + y_2$ означает, что $x + (\lambda_1 - \lambda_2)e + y_1 = y_2$ в случае $\lambda_1 > \lambda_2$, или $x + y_1 = (\lambda_2 - \lambda_1)e + y_2$ в противном случае. Поэтому для всякого $x \in Lin(e, Y)$ достаточно рассмотреть случаи $x + \lambda e + y_1 = y_2$ или $x + y_1 = \lambda e + y_2$ ($\lambda \geqslant 0$ и $y_1, y_2 \in Y$). Определим L на $Q = Lin(e, Y)$ следующим образом:

1) Если $x_1 \in Q$: $x_1 + y_{12} = \lambda_1 e + y_{11}$ для некоторых y_{11}, $y_{12} \in Y$ и $\lambda_1 \geqslant 0$, тогда

$$L(x_1) := \lambda_1 L(e) + \ell(y_{11}) - \ell(y_{12}).$$

2) Если существует $x_2 \in Q$: $x_2 + \lambda_2 e + y_{22} = y_{21}$ для некоторого y_{21}, $y_{22} \in Y$ и $\lambda_2 \geqslant 0$, тогда

$$L(x_2) := \ell(y_{21}) - \ell(y_{22}) - \lambda_2 L(e).$$

Необходимо доказать существование такого числа $L(e)$, что $L(x_i) \leqslant p(x_i)$ для любого $i = 1, 2$, или

$$\lambda_1 L(e) + \ell(y_{11}) - \ell(y_{12}) \leqslant p(x_1), \quad -\lambda_2 L(e) + \ell(y_{21}) - \ell(y_{22}) \leqslant p(x_2). \quad (5.57)$$

Случай $\lambda_i = 0$ для некоторого $i = 1, 2$ очевиден: $x_i \in Y$. Таким образом, без ограничения общности будем считать, что $\lambda_i > 0$ для $i = 1, 2$ и возможно переписать условие (5.57) следующим образом:

$$L(e) \leqslant p\left(\frac{x_1}{\lambda_1}\right) + \ell\left(\frac{y_{12}}{\lambda_1}\right) - \ell\left(\frac{y_{11}}{\lambda_1}\right), \quad L(e) \geqslant \ell\left(\frac{y_{21}}{\lambda_2}\right) - \ell\left(\frac{y_{22}}{\lambda_2}\right) - p\left(\frac{x_2}{\lambda_2}\right).$$

для всех возможных $y_{11}, y_{12}, y_{21}, y_{22} \in Y$ и $\lambda_i > 0$ ($i = 1, 2$).

Давайте проверим полезное неравенство:

$$\ell\left(\frac{y_{21}}{\lambda_2}\right) - \ell\left(\frac{y_{22}}{\lambda_2}\right) - p\left(\frac{x_2}{\lambda_2}\right) \leqslant p\left(\frac{x_1}{\lambda_1}\right) + \ell\left(\frac{y_{12}}{\lambda_1}\right) - \ell\left(\frac{y_{11}}{\lambda_1}\right). \quad (5.58)$$

для всех возможных $y_{11}, y_{12}, y_{21}, y_{22} \in Y$ и $\lambda_i > 0$ $(i = 1, 2)$.

Действительно, в силу закона сокращения на Q из равенств

$$\frac{x_1}{\lambda_1} + \frac{y_{12}}{\lambda_1} = \frac{y_{11}}{\lambda_1} + e \quad \text{и} \quad \frac{x_2}{\lambda_2} + \frac{y_{22}}{\lambda_2} + e = \frac{y_{21}}{\lambda_2}$$

получаем

$$\frac{x_1}{\lambda_1} + \frac{x_2}{\lambda_2} + \frac{y_{12}}{\lambda_1} + \frac{y_{22}}{\lambda_2} = \frac{y_{11}}{\lambda_1} + \frac{y_{21}}{\lambda_2},$$

то есть $\frac{x_1}{\lambda_1} + \frac{x_2}{\lambda_2} \in Y$, так что Y — подконус Q (см. определение 5.7.15). Далее,

$$p\left(\frac{x_1}{\lambda_1}\right) + \ell\left(\frac{y_{12}}{\lambda_1}\right) - \ell\left(\frac{y_{11}}{\lambda_1}\right) - \left(\ell\left(\frac{y_{21}}{\lambda_2}\right) - \ell\left(\frac{y_{22}}{\lambda_2}\right) - p\left(\frac{x_2}{\lambda_2}\right)\right) =$$

$$= p\left(\frac{x_1}{\lambda_1}\right) + p\left(\frac{x_2}{\lambda_2}\right) + \ell\left(\frac{y_{12}}{\lambda_1} + \frac{y_{22}}{\lambda_2}\right) - \ell\left(\frac{y_{11}}{\lambda_1} + \frac{y_{21}}{\lambda_2}\right) =$$

$$= p\left(\frac{x_1}{\lambda_1}\right) + p\left(\frac{x_2}{\lambda_2}\right) + \ell\left(\frac{y_{12}}{\lambda_1} + \frac{y_{22}}{\lambda_2}\right) - \ell\left(\frac{x_1}{\lambda_1} + \frac{x_2}{\lambda_2} + \frac{y_{12}}{\lambda_1} + \frac{y_{22}}{\lambda_2}\right) =$$

$$= p\left(\frac{x_1}{\lambda_1}\right) + p\left(\frac{x_2}{\lambda_2}\right) - \ell\left(\frac{x_1}{\lambda_1} + \frac{x_2}{\lambda_2}\right) \geqslant p\left(\frac{x_1}{\lambda_1} + \frac{x_2}{\lambda_2}\right) - \ell\left(\frac{x_1}{\lambda_1} + \frac{x_2}{\lambda_2}\right) \geqslant 0$$

с учетом $\ell(y) \leqslant p(y)$ для любого $y \in Y$.

Для удобства введем следующее соглашение. Если для некоторого $\lambda_1 > 0$ и $y_{11}, y_{12} \in Y$ не существует $x_1 \in Q$: $x_1 + y_{12} = \lambda_1 e + y_{11}$, то полагаем для x_1 следующее:

$$p\left(\frac{x_1}{\lambda_1}\right) + \ell\left(\frac{y_{12}}{\lambda_1}\right) - \ell\left(\frac{y_{11}}{\lambda_1}\right) := +\infty. \tag{5.59}$$

Если для некоторого $\lambda_2 > 0$ и $y_{21}, y_{22} \in Y$ не существует $x_2 \in Q$: $x_2 + y_{22} + \lambda_2 e = y_{21}$, то полагаем для x_2:

$$\ell\left(\frac{y_{21}}{\lambda_2}\right) - \ell\left(\frac{y_{22}}{\lambda_2}\right) - p\left(\frac{x_2}{\lambda_2}\right) := -\infty. \tag{5.60}$$

В силу (5.58) $\alpha_e \leqslant \beta_e$, где

$$\alpha_e := \sup_{\lambda_2 > 0,\, y_{21}, y_{22} \in Y} \left\{\ell\left(\frac{y_{21}}{\lambda_2}\right) - \ell\left(\frac{y_{22}}{\lambda_2}\right) - p\left(\frac{x_2}{\lambda_2}\right)\right\}$$

и

$$\beta_e := \inf_{\lambda_1, > 0,\, y_{11}, y_{12} \in Y} \left\{p\left(\frac{x_1}{\lambda_1}\right) + \ell\left(\frac{y_{12}}{\lambda_1}\right) - \ell\left(\frac{y_{11}}{\lambda_1}\right)\right\}.$$

Можно выбрать $L(e) \in [\alpha_e; \beta_e]$ и тогда будут выполняться условия (5.57).

2) Очевидно, что ℓ может быть аналогичным образом распространена на последовательность подконусов

$$Lin(e_1, Y) \subset Lin(e_2, Y_1) \subset \ldots \subset Lin(e_n, Y_{n-1}) \subset \ldots,$$

где $Y_k = Lin(e_k, Y_{k-1})$, $Y = Y_0$, $Y_1 \subset Y_2 \subset \ldots$

Введем следующий порядок: $(Z, L_Z) \leqslant (Z_1, L_{Z_1})$, если $Z \subset Z_1$ и линейный функционал L_{Z_1} — расширение L_Z. По лемме Цорна каждое упорядоченное множество имеет мажоранту: объединение. Следовательно, существует максимальный элемент (Y_m, L_m). Согласно п. 1), $Y_m = Q$. Это завершает доказательство. □

Перейдем к аналогу хорошо известной леммы об опорном функционале в нормированных конусах. Введём аналог сопряжённого пространства в нормированных конусах. Как и в классическом случае будем говорить, что линейный функционал ℓ *ограничен* на Q, если для некоторого $C > 0$ верно $\ell(x) \leqslant C\|x\|$ $\forall x \in Q$.

Отметим (см., например, [12]), что такой функционал может быть не ограничен снизу (т.е. возможно $\lim\limits_k \ell(x_k) = -\infty$ для ограниченной по норме последовательности $\{x_k\}_{k=1}^{+\infty}$). Ясно, что множество всех ограниченных линейных функционалов на Q будет выпуклым конусом, если в нём стандартно ввести операцию сложения и умножения на неотрицательный скаляр. Выделим класс линейных ограниченных функционалов, принимающих неотрицательные значения в ненулевых точках Q.

Определение 5.7.17. *Сопряженным конусом* к нормированному конусу Q назовём множество Q^* всех ограниченных линейных функционалов $\ell : Q \to \mathbb{R}$ таких, что $\ell(x_0) \geqslant 0$ для некоторого ненулевого элемента $x_0 \in Q$.

В Q^* можно естественным образом ввести следующую полунорму:

$$\|\ell\|_* := \sup_{x \neq 0} \left\{ \frac{\ell(x)}{\|x\|} \right\}. \tag{5.61}$$

Из теоремы 5.7.16 вытекает аналог известной леммы об опорном функционале в классе выпуклых нормированных конусов.

Следствие 5.7.18. *Пусть Q — нормированный конус с законом сокращения. Тогда для всякого фиксированного $x_0 \in Q\backslash\{0\}$ существует такой $\ell \in Q^* \setminus \{0\}$, что $\|\ell\|_* = 1$ и $\ell(x_0) = \|x_0\|$.*

Доказательство. Легко видеть, что $Y = \{x \in Q \,|\, x+\mu x_0 = \lambda x_0, \lambda \text{ и } \mu \geqslant 0\}$ — подконус Q. Так как Q имеет закон сокращения, то $x + \mu x_0 = \lambda x_0$ означает, что $x + (\mu - \lambda)x_0 = 0$ для $\mu > \lambda$ или $x = (\lambda - \mu)x_0$ для $\lambda \geqslant \mu$. Положим $\ell(x) := (\lambda - \mu)\|x_0\|$. Отметим, что если $\lambda < \mu$, то $\ell(x) < 0 \leqslant \|x\|$. Остаётся лишь применить теорему 5.7.16. □

Замечание 5.7.19. Аналогичный результат известен в специальных классах нормированных конусов Q для неотрицательных линейных функционалов $f : Q \to \mathbb{R}^+$ (см. [165], теорема 2.14). Однако при этом, вообще говоря, равенство $f(x_0) = \|x_0\|$ невозможно (см. замечание после теоремы 2.14 в [165]).

Замечание 5.7.20. Аналогичный результат можно получить и для полунормированных конусов с полунормой $\|\cdot\|$. Для этого (5.61) нужно видоизменить следующим образом

$$\|\ell\|_* := \sup_{x\ \|x\|\neq 0} \left\{ \frac{\ell(x)}{\|x\|} \right\}.$$

С такой оговоркой утверждение следствия 5.7.18 будет верно для всякого $x \in Q$ такого, что $\|x\| \neq 0$.

5.7.5 Отделимые нормированные конусы: определение и примеры

В данном пункте работы мы введём базовое понятие *отделимого нормированного конуса* и рассмотрим примеры таких конусов. В частности, показано что выпуклые конусы во всяком нормированном пространстве, а также в пространстве с несимметричной нормой есть отделимые нормированные конусы.

Определение 5.7.21. Будем говорить, что нормированный конус Q с законом сокращения отделим, если для всяких различных элементов x_1 и x_2 из Q существует такой линейный функционал $\ell : Q \to \mathbb{R}$, что $\|\ell\|_* = 1$ и

$$\ell(x_1) \neq \ell(x_2), \text{ причём } \ell(x_1) > 0 \text{ или } \ell(x_2) > 0. \tag{5.62}$$

Замечание 5.7.22. Отметим, что не всякий нормированный конус с законом сокращения отделим. Например, это относится к набору числовых пар:

$$Q = \{(0,0)\} \cup \{(a,b) \,|\, a > 0 \ \text{ и } \ b \in \mathbb{R}\}$$

с нормой $||(a,b)|| = a$. Нетрудно видеть, что любой линейный ограниченный функционал $\ell : Q \to \mathbb{R}$ имеет вид $\ell((a,b)) = \lambda a$ для некоторого $\lambda \geqslant 0$. Очевидно, что такие функционалы не разделяют точки Q и поэтому Q не удовлетворяет (5.62).

Тем не менее, можно привести классы примеров нормированных конусов, удовлетворяющих определению 5.7.21. Начнём с практически очевидного примера.

Пример 5.7.23. Всякое нормированное пространство E есть отделимый нормированный конус. Действительно, для любых $x_1, x_2 \in E$, $x_1 \neq x_2$, существует такой $\ell \in E^*$, что $\ell(x_1) \neq \ell(x_2)$. Если $\ell(x_1) > 0$ или $\ell(x_2) > 0$, то все доказано. Если же $\ell(x_1) \leqslant 0$ и $\ell(x_2) \leqslant 0$, то можно рассмотреть $-\ell \in E^*$.

К предыдущему примеру естественно примыкает следующий.

Пример 5.7.24. Хорошо известно [161], что выпуклый конус Q с нормой линейно, инъективно изометрично вложен в нормированное пространство E тогда и только тогда, когда в Q существует такая однородная и инвариантная относительно сдвигов метрика $d : Q \times Q \longrightarrow \mathbb{R}$, что $d(0,x) = \|x\|$ для всякого $x \in Q$. Всякий выпуклый конус Q с такой метрикой мы будем называть *линейным нормированным конусом* (ЛНК). Ясно, $\ell \in E^*$ тогда и только тогда, когда $-\ell \in E^*$. Поэтому всякий ЛНК будет отделимым нормированным конусом.

Перейдем к более нетривиальным классам примеров. Существуют примеры отделимых нормированных конусов, которые не могут быть линейно инъективно изометрично вложены ни в какое банахово пространство.

Пример 5.7.25. Докажем, что *всякое пространство E с несимметричной нормой $||\cdot|$ есть отделимый нормированный конус.* В указанном случае для ограниченного функционала $\ell : Q \to \mathbb{R}$ функционал $-\ell$ необязательно ограничен, поэтому рассуждения не столь тривиальны. Пусть $x_1 \neq x_2$ из E.

1) Если $||x_1| \neq ||x_2|$, то для определенности можно полагать $||x_2| > ||x_1|$ и по следствию 5.7.18 рассмотреть $\ell \in E^*$ такой, что $\ell(x_1) = ||x_1|$. Тогда

$$\ell(x_2) \leqslant ||x_2| < ||x_1| = \ell(x_1) > 0.$$

2) Пусть теперь $||x_1| = ||x_2|$. Возьмем опять $\ell \in E^*$: $\ell(x_1) = ||x_1|$. Если $\ell(x_2) \neq ||x_1| = ||x_2|$, то все доказано. Если же $\ell(x_1) = \ell(x_2) = ||x_1| = ||x_2| > 0$, то выберем $\ell_1 \in E^*$: $\ell_1(x_1 - x_2) = ||x_1 - x_2| > 0$ ввиду $x_1 \neq x_2$. Это означает, что $\ell_1(x_1) = \ell_1(x_2) + ||x_1 - x_2|$. Поэтому при всяком $C > 0$ верно $\ell_1 + C\ell \in E^*$ и

$$(\ell_1 + C\ell)(x_1) = \ell_1(x_2) + ||x_1 - x_2| + C||x_1| =$$

$$= \ell_1(x_2) + C||x_2| + ||x_1 - x_2| > (\ell_1 + C\ell)(x_2).$$

Поскольку $||x_1| > 0$, то можно выбрать такое $C_0 > 0$, что $(\ell_1 + C_0\ell)(x_1) > 0$. Поэтому при всех $x_1 \neq x_2$ из E существует такой $\ell \in E^*$, что $\ell(x_1) \neq \ell(x_2)$ и $\ell(x_1) > 0$, либо $\ell(x_2) > 0$, что и требовалось доказать.

Ясно, что любой конус в линейном пространстве с несимметричной нормой будет отделимым нормированным конусом.

Перейдём к примерам классов нормированных конусов, которые не являются линейными пространствами.

Пример 5.7.26. Отделимым нормированным конусом будет всякий нормированный конус, в котором линейные неотрицательные ограниченные функционалы разделяют точки. Отметим, что такие конусы были исследованы, в частности, в работах [100, 162].

Возможны также другие типы примеров отделимых нормированных конусов, не вложенных линейно инъективно изометрично в нормированное пространство (как с обычной, так и с несимметричной нормой). В частности, такой пример предложен в ([171], пример 8). Приведём ещё 1 пример, показывающий, что норма, заданная на выпуклом конусе Q, может не быть полунормой ни в каком линейном пространстве E, в которое линейно инъективно вложен Q.

Пример 5.7.27. Рассмотрим набор числовых пар:

$$Q = \{(0,0)\} \cup \{(a,b) \,|\, a > 0 \text{ и } b > 0\}$$

с нормой $||(a,b)|| = \max\left\{a, \frac{b^2}{a}\right\}$ при $a \neq 0$ и $||(0,0)|| = 0$.

Ясно, что функционалы $\ell_1((a,b)) = a$ и $\ell_2((a,b)) = a+b$ разделяют все ненулевые точки Q. Кроме того, они неотрицательны, ограничены на единичном шаре $\{(a,b)\,|\,||(a,b)|| \leqslant 1\} \subset Q$ и поэтому удовлетворяют определению 5.7.21.

Отметим, что к примеру при $(a,b) = (1,7)$ и $(c,d) = (7,1)$ верно

$$||(a,b)|| + ||(a,b) + (c,d)|| < ||(c,d)||,$$

что указывает на невозможность линейного инъективного изометричного вложения данного нормированного конуса ни в какое банахово пространство.

Теперь мы обоснуем возможность рассматривать всякий отделимый нормированный конус Q как метрическое пространство.

Согласно определению 5.7.21 во всяком отделимом нормированном конусе Q для любых различных $x_1, x_2 \in Q$ существует такой линейный функционал $\ell : Q \to \mathbb{R}$, что $\ell(x_1) \neq \ell(x_2)$, причем $\ell(x_1) > 0$ или $\ell(x_2) > 0$, а также $\ell(x) \leqslant ||x||\ \forall x \in Q$. Каждому такому функционалу ℓ можно поставить в соответствие выпуклый функционал:

$$p_\ell(x) = \max\{0, \ell(x)\}.$$

Обозначим через $\widehat{Q}^*_{sub}$ выпуклый конус, содержащий все функционалы вида (5.7.5) со следующими операциями сложения и умножения на неотрицательный скаляр:

$$[p_{\ell_1} \oplus p_{\ell_2}](\cdot) := p_{\ell_1+\ell_2}(\cdot); \quad p_{\lambda\ell}(\cdot) := \lambda \cdot p(\cdot)$$

для произвольных $\ell, \ell_1, \ell_2 \in Q^*$, $\lambda \geqslant 0$. На $\widehat{Q}^*_{sub}$ естественно можно ввести норму:

$$||p||_* := \sup_{x \in Q\setminus\{0\}} \left\{ \frac{p(x)}{||x||} \right\} \ \forall p \in \widehat{Q}^*_{sub}. \tag{5.63}$$

Определение 5.7.28. Выпуклый конус $\widehat{Q}^*_{sub}$ с нормой (5.63) будем называть *субсопряжённым конусом* к Q.

Из определения 5.7.21 следует, что функционалы вида (5.7.5) отделяют точки всякого отделимого нормированного конуса Q. Пусть $T \subset \widehat{Q}^*_{sub}$ — набор функционалов нормы 1, разделяющих точки Q. Поэтому на Q можно ввести метрику

$$d(x,y) := \sup_{||\ell||_*=1} |\max\{0,\ell(x)\} - \max\{0,\ell(y)\}| = \sup_{||p_\ell||_*=1} |p_\ell(x) - p_\ell(y)|. \tag{5.64}$$

Заметим, что метрика (5.64) однородна, т.е. для всяких $x, y \in Q$ и $\lambda \geqslant 0$:

$$d(\lambda x, \lambda y) = \lambda \sup_{||p_\ell||_*=1} |p_\ell(x) - p_\ell(y)| = \lambda d(x, y).$$

Таким образом, верно следующее утверждение.

Лемма 5.7.29. *Во всяком отделимом нормированном конусе существует такая однородная метрика $d : Q \times Q \to \mathbb{R}$, что $d(0, x) = ||x||$.*

Предыдущее утверждение позволяет рассматривать Q как метрическое пространство (Q, d). Отметим следующую оценку для такой метрики в классе линейных пространств E с несимметричной нормой $\|\cdot|$. Всюду далее $||x||_{sym} = \max\{||x|, ||-x|\}$ — согласованная с $||\cdot|$ обычная симметричная норма в E.

Лемма 5.7.30. *Если E — линейное пространство с несимметричной нормой $||\cdot|$. Тогда для любых $x, y \in E$ справедливо неравенство*

$$d(x, y) \leqslant ||x - y||_{sym}. \tag{5.65}$$

Доказательство. Ясно, что $d(x, y) \leqslant \sup\limits_{\|\ell\|_*=1} \max\{|\ell(x)-\ell(y)|, \ell(x), \ell(y)\}$, поскольку при $\ell(x) < 0$ и $\ell(y) < 0$ $p_\ell(x) = p_\ell(y) = 0$, а $|\ell(x) - \ell(y)| \geqslant 0$.

Очевидно, что $\ell(x) \leqslant ||x|$ $\forall x \in E$ означает, что $\ell(x) \leqslant ||x||_{sym}$ при всех $x \in E$, то есть $\ell \in E^*_{sym}$, где $E = (E_{sym}, ||\cdot||_{sym})$ — обычное нормированное пространство. Тогда $|\ell(x)| \leqslant ||x||_{sym}$ и $|\ell(y)| \leqslant ||y||_{sym}$ для любых $x, y \in E$, откуда

$$|\ell(x) - \ell(y)| = |\ell(x - y)| \leqslant ||x - y||_{sym}.$$

При $\ell(x) \geqslant 0$ и $\ell(y) < 0$ верно $\ell(x) \leqslant \ell(x) - \ell(y) \leqslant |\ell(x) - \ell(y)| \leqslant ||x - y||_{sym}$. Аналогичную оценку можно получить при $\ell(x) < 0$ и $\ell(y) \geqslant 0$. Поэтому верно неравенство (5.65). □

Замечание 5.7.31. Отметим, что равенство в (5.65) может и не выполняться. Например, если $y = -x \neq 0$, то для всякого $\ell \in Q^*$ верно $\ell(y) = -\ell(x)$. Поэтому $p_\ell(x) = 0$ или $p_\ell(y) = 0$ и это означает, что для всякого $\ell \in Q^*$, $\|\ell\|_* = 1$ верны неравенства $|p_\ell(x) - p_\ell(y)| < |\ell(x) - \ell(y)| = |\ell(x - y)| \leqslant \|x - y\|_{sym}$, т.е. в данном случае $d(x, y) < \|x - y\|_{sym}$.

Рассмотрим ещё некоторый нетривиальный пример класса отделимых нормированных конусов.

Определение 5.7.32. Множество Q называют строгим конусом, если выполнено следующее свойство:

$$x + y = 0 \Longrightarrow x = y = 0 \text{ для любых } x, y \in Q. \tag{5.66}$$

Заметим, что строгие выпуклые конусы были ранее рассмотрены, к примеру, в работе [165]. Во всяком строгом выпуклом конусе Q возможно ввести следующий частичный порядок [165]:

$$x \preceq y \text{ если } y = x + z \text{ для некоторого } z \in Q.$$

В [171] введено понятие строгого выпуклого нормированного конуса, отличительная особенность которого — следующее свойство *порядковой отделимости* (5.67).

Определение 5.7.33. Будем говорить, что абстрактный выпуклый конус Q называется *строгим выпуклым нормированным конусом* (СВНК), если Q — строгий выпуклый нормированный конус с законом сокращения такой, что

(i) для всяких $x, y \in Q$ и произвольных неотрицательных чисел $\alpha < 1 < \beta$

$$\alpha x \preceq y \preceq \beta x \Longrightarrow y = x; \tag{5.67}$$

(ii) для всяких $x, y \in Q$ из $x \preceq y$ следует, что $\|x\| \leqslant \|y\|$.

Теперь рассмотрим некоторые примеры строгих нормированных конусов.

Пример 5.7.34. Пусть $Q = \mathbb{R}_+$ — набор неотрицательных чисел с обычной операцией сложения и умножения на скаляр. Для любых $a, b \in \mathbb{R}_+$ $a + b = 0$ означает $a = b = 0$, и

$$a \leqslant b \Longleftrightarrow b = a + c \text{ для некоторого } c \geqslant 0.$$

Ясно, что $\alpha b \preceq a \preceq \beta b$ для любого $\alpha < 1 < \beta$ означает $b \leqslant a \leqslant b$, то есть $a = b$.

Пример 5.7.35. Пусть $Q = m_+$ — набор последовательностей $a = \{a_n\}_{n=1}^{\infty}$, где $a_n \geqslant 0$ для любого $n \in \mathbb{N}$, 0 — нулевая последовательность. Операции сложения и умножения на скаляр в m_+ задаются стандартно: $a+b := \{a_n+b_n\}_{n=1}^{\infty}$ и $\lambda \cdot a := \{\lambda \cdot a_n\}_{n=1}^{\infty}$ для любого $\lambda \geqslant 0$ и $a = \{a_n\}_{n=1}^{\infty}$, $b = \{b_n\}_{n=1}^{\infty}$ из m_+. Заметим, что $a \preceq b$ тогда и только тогда, когда $a_n \leqslant b_n$ для любого $n \in \mathbb{N}$.

Пример 5.7.36. Пусть $Q = F_+$ — набор неотрицательных ограниченных вещественных функций $f : [0;1] \to \mathbb{R}_+$ с обычными операциями сложения и умножения на скаляр. В этом случае для любых $f, g \in F_+$ верно $f \preceq g$ тогда и только тогда, когда $f(t) \leqslant g(t)$ для любого $t \in [0;1]$.

Пример 5.7.37. Пусть $Q = K_+$ — набор отрезков $A = [a;b] \subset \mathbb{R}^+$ со сложением по Минковскому и обычным скалярным умножением. В этом случае мы можем рассматривать следующий порядок: для любых $A, B \in F_+$ $A \preceq B$ тогда и только тогда, когда существует такое $C \in K_+$, что $B = A + C$.

Теперь мы переходим к аналогу теоремы Хана–Банаха об отделимости точек в (СВНК). Этот результат важен для доказательства условий метризуемости нормированного конуса.

Теорема 5.7.38. *Пусть Q — (СВНК). Тогда для любых $e_1 \neq e_2$ $(e_1, e_2 \in Q)$ существует такой функционал $\ell \in Q^* \backslash \{0\}$, что $\ell(e_1) \neq \ell(e_2)$ и при этом $\ell(e_1) > 0$ или $\ell(e_2) > 0$.*

Доказательство. 1) Если $\|e_1\| \neq \|e_2\|$ (например, $e_1 = 0$ или $e_2 = 0$), тогда для любого $i = 1, 2$ по следствию 5.7.18 существует такое $\ell_i \in Q^* \backslash \{0\}$, что $\ell_i(e_i) = \|e_i\|$. Если $\|e_1\| < \|e_2\|$, тогда по следствию 5.7.18 мы имеем $\ell_2(e_1) \leqslant \|e_1\| < \|e_2\| = \ell_2(e_2)$ и $\ell_2(e_1) \neq \ell_2(e_2) > 0$. Случай $\|e_2\| < \|e_1\|$ рассматривается аналогично.

2) Теперь предположим, что $\|e_1\| = \|e_2\|$, $e_1 \neq 0$ и $e_2 \neq 0$. Положим $Y := \{\lambda e_1 | \lambda \geqslant 0\}$.

Заметим, что Y — подконус Q. Действительно, в силу закона сокращения для любого $\lambda \geqslant \mu \geqslant 0$ из $x + \mu e_1 = \lambda e_1$ имеем $x = (\lambda - \mu)e_1$. В случае $\lambda < \mu$ невозможно ввиду строгости Q ($x + (\mu - \lambda)e_1 = 0$ означает $x = e_1 = 0$).

Для всякого $x = \lambda e_1 \in Y$ можно определить такое $\ell \in Q^*$, что $\ell(x) := \lambda \|e_1\|$. Ясно, что для $e_2 \in Y \backslash \{e_1\}$ имеем $\ell(e_2) \neq \|e_1\| = \ell(e_1) > 0$.

Если $e_2 \notin Y$ тогда аналогично доказательству теоремы 5.7.16 рассмотрим множество $Y_1 = Lin(e_2, Y)$. В силу закона сокращения на Q для любого $y_1 = \mu_1 e_1$ и $y_2 = \mu_2 e_1 \in Y$ из $x + \lambda_2 e_2 + \mu_2 e_1 = \lambda_1 e_2 + \mu_1 e_1$ имеем

$$x = (\lambda_1 - \lambda_2)e_2 + (\mu_1 - \mu_2)e_1, \text{ для } \lambda_1 \geqslant \lambda_2 \text{ и } \mu_1 \geqslant \mu_2,$$

$$x + (\lambda_2 - \lambda_1)e_2 = (\mu_1 - \mu_2)e_1, \text{ для } \lambda_1 < \lambda_2 \text{ и } \mu_1 \geqslant \mu_2,$$

$$x + (\mu_2 - \mu_1)e_1 = (\lambda_1 - \lambda_2)e_2, \text{ для } \lambda_1 \geqslant \lambda_2 \text{ и } \mu_1 < \mu_2.$$

Равенство $x + (\mu_2 - \mu_1)e_1 + (\lambda_2 - \lambda_1)e_2 = 0$ означает $x = (\mu_2 - \mu_1)e_1 = (\lambda_2 - \lambda_1)e_2 = 0$ ввиду (5.66). Поэтому достаточно рассмотреть только 3 типа элементов
$x_1, x_2, x_3 \in Y_1$:

а) $x_1 = \lambda_1 e_2 + \alpha_1\lambda_1 e_1$ для некоторого $\alpha_1, \lambda_1 \geqslant 0$;

б) $x_2 + \lambda_2 e_2 = \alpha_2\lambda_2 e_1$ в случае существования $x_2 \in Q$ для некоторого $\alpha_2, \lambda_2 \geqslant 0$;

в) $x_3 + \alpha_3\lambda_3 e_1 = \lambda_3 e_2$ в случае существования $x_3 \in Q$ для некоторого $\alpha_3, \lambda_3 \geqslant 0$.

Для $x_1, x_2, x_3 \in Q$ условия (5.57) на $\ell(e_2)$ могут быть записаны следующим образом:

$$\lambda_1\ell(e_2) + \alpha_1\lambda_1\ell(e_1) \leqslant p(x_1),$$

$$-\lambda_2\ell(e_2) + \alpha_2\lambda_2\ell(e_1) \leqslant p(x_2), \;\; \lambda_3\ell(e_2) - \alpha_3\lambda_3\ell(e_1) \leqslant p(x_3),$$

или, что эквивалентно, для всех возможных $\alpha_i > 0$ и $\lambda_i > 0$ $(i = 1, 2, 3)$

$$\ell(e_2) \leqslant p\left(\frac{x_1}{\lambda_1}\right) - \alpha_1\ell(e_1), \;\; \ell(e_2) \geqslant \alpha_2\ell(e_1) - p\left(\frac{x_2}{\lambda_2}\right),$$

$$\ell(e_2) \leqslant p\left(\frac{x_3}{\lambda_3}\right) + \alpha_3\ell(e_1).$$

В соответствии с доказательством теоремы 5.7.16 можно выбрать $\ell(e_2) \in [\alpha_{e_2}; \beta_{e_2}]$ (мы используем соглашения (5.59) и (5.60)), где для всевозможных $\alpha_i > 0$ и $\lambda_i > 0$ $(i = 1, 2, 3)$

$$\alpha_{e_2} = \sup_{\alpha_2, \lambda_2 > 0} \left\{\alpha_2\ell(e_1) - p\left(\frac{x_2}{\lambda_2}\right)\right\}$$

и

$$\beta_{e_2} = \inf_{\alpha_1, \alpha_3, \lambda_1, \lambda_3 > 0} \left\{p\left(\frac{x_1}{\lambda_1}\right) - \alpha_1\ell(e_1);\; p\left(\frac{x_3}{\lambda_3}\right) + \alpha_3\ell(e_1)\right\}.$$

В силу соглашений (5.59) и (5.60) можно считать, не уменьшая общности рассуждений, что x_2 и x_3 существуют в Q для некоторых $\alpha_i > 0$ и $\lambda_i > 0$ $(i = 2, 3)$. Так как

$$\alpha_2 e_1 = \frac{x_2}{\lambda_2} + e_2 \;\text{ и }\; \alpha_3 e_1 + \frac{x_3}{\lambda_3} = e_2,$$

то $\alpha_3 e_1 \preceq e_2 \preceq \alpha_2 e_1$ для некоторых $\alpha_2 \geqslant \alpha_3 \geqslant 0$.

Если $\inf\{\alpha_2 - \alpha_3\} = 0$, тогда из $\|e_1\| = \|e_2\|$ и $\alpha_3\|e_1\| \leqslant \|e_2\| \leqslant \alpha_2\|e_1\|$ имеем $\sup \alpha_3 = \inf \alpha_2 = 1$. Тогда ввиду (5.67) $e_1 = e_2$, что невозможно.

Пусть теперь $\inf\{\alpha_2 - \alpha_3\} = \delta > 0$. Поскольку Q имеет закон сокращения, то

$$\frac{x_2}{\lambda_2} + \frac{x_3}{\lambda_3} + \alpha_3 e_1 = \alpha_2 e_1 \quad \text{и} \quad \frac{x_2}{\lambda_2} + \frac{x_3}{\lambda_3} = (\alpha_2 - \alpha_3)e_1 \in Y.$$

Следовательно,

$$\left\|\frac{x_3}{\lambda_3}\right\| + \alpha_3 \ell(e_1) - \left(\alpha_2 \ell(e_1) - \left\|\frac{x_2}{\lambda_2}\right\|\right) =$$

$$= \left\|\frac{x_2}{\lambda_2}\right\| + \left\|\frac{x_3}{\lambda_3}\right\| + \alpha_3 \ell(e_1) - \ell\left((\alpha_2 - \alpha_3)e_1 + \alpha_3 e_1\right) =$$

$$= \left\|\frac{x_2}{\lambda_2}\right\| + \left\|\frac{x_3}{\lambda_3}\right\| - \ell\left((\alpha_2 - \alpha_3)e_1\right) \geqslant \left\|\frac{x_2}{\lambda_2} + \frac{x_3}{\lambda_3}\right\| - \frac{1}{2}\|(\alpha_2 - \alpha_3)e_1\| =$$

$$= \|(\alpha_2 - \alpha_3)e_1\| - \frac{1}{2}\|(\alpha_2 - \alpha_3)e_1\| = \frac{1}{2}\|(\alpha_2 - \alpha_3)e_1\| \geqslant \frac{1}{2}\delta\|e_1\| > 0 \quad \text{и}$$

$$\left\|\frac{x_1}{\lambda_1}\right\| - \alpha_1 \ell(e_1) - \left(\alpha_2 \ell(e_1) - \left\|\frac{x_2}{\lambda_2}\right\|\right) =$$

$$= \left\|\frac{x_1}{\lambda_1}\right\| + \left\|\frac{x_2}{\lambda_2}\right\| - \frac{1}{2}\ell\left(\alpha_1 e_1 + \alpha_2 e_1\right) =$$

$$= \|\alpha_1 e_1 + e_2\| + \left\|\frac{x_2}{\lambda_2}\right\| - \frac{1}{2}\|\alpha_1 e_1 + \alpha_2 e_1\| \geqslant \frac{1}{2}\|\alpha_1 e_1 + e_2\| + \frac{1}{2}\left\|\frac{x_2}{\lambda_2}\right\| +$$

$$+ \frac{1}{2}\left\|\alpha_1 e_1 + e_2 + \frac{x_2}{\lambda_2}\right\| - \frac{1}{2}\|\alpha_1 e_1 + \alpha_2 e_1\| =$$

$$= \frac{1}{2}\|\alpha_1 e_1 + e_2\| + \frac{1}{2}\left\|\frac{x_2}{\lambda_2}\right\| \geqslant \frac{1}{2}\|e_2\| > 0$$

ввиду $\|x\| \leqslant \|y\|$ для $x = e_2 \preceq y = \alpha_1 e_1 + e_2$ (см. определение 5.7.33).

Поэтому $\beta_{e_2} \geqslant \alpha_{e_2} + \min\left\{\frac{1}{2}\delta\|e_1\|; \frac{1}{2}\|e_2\|\right\}$, то есть $\beta_{e_2} > \alpha_{e_2}$, и возможно выбрать $\ell \in Q^*$ так, чтобы $\ell(e_2) \neq \ell(e_1) = \frac{1}{2}\|e_1\| > 0$. □

В заключении настоящего раздела отметим, что предыдущий результат обобщён в [173] на следующий класс выпуклых конусов с нормой.

Определение 5.7.39. Будем называть что абстрактный выпуклый конус Q выпуклым упорядоченным нормированным конусом (ВУНК), если Q есть строгий выпуклый порядково отделимый нормированный конус с законом сокращения и для произвольного $x \in Q$:

$$x \neq 0 \Rightarrow inf\{\|y\| | x \preceq y\} > 0. \tag{5.68}$$

Очевидно, (5.68) следует из определения (СВНК). Следовательно, всякий (СВНК) будет также и (ВУНК). Однако существуют (ВУНК), который не является (СВНК). Рассмотрим пример такого конуса.

Пример 5.7.40. Пусть Q — множество пар неотрицательных чисел (a, b):

$$Q = \{(0,0)\} \cup \{(a,b) \,|\, a > 0 \text{ и } b > 0\}.$$

Введем норму в Q следующим образом:

$$||(a,b)|| = \max\left\{a, \frac{b^2}{a}\right\} \text{ для } a \neq 0 \text{ и } ||(0,0)|| = 0.$$

Ясно, что $||(a,b)|| = p_W((a,b))$, где $p_W(\cdot)$ — функционал Минковского множества W, ограниченный параболой $a = b^2$ и прямой линией $a = 1$. Поэтому $||\cdot||$ — выпуклый функционал на Q. Покажем, что $||\cdot||$ может не быть полунормой в линейном пространстве $E \supset \varphi(Q)$ для всякого линейного инъективного вложения $\varphi : Q \to E$. Действительно, для пар $\left(\frac{1}{8}, \frac{7}{8}\right) + \left(\frac{7}{8}, \frac{1}{8}\right) = (1,1)$ имеем:

$$\left\|\left(\frac{1}{8}, \frac{7}{8}\right)\right\| = \frac{49}{8}, \ \left\|\left(\frac{7}{8}, \frac{1}{8}\right)\right\| = \frac{7}{8}, \ ||(1,1)|| = 1,$$

$$||(1,1)|| + \left\|\left(\frac{7}{8}, \frac{1}{8}\right)\right\| < \left\|\left(\frac{1}{8}, \frac{7}{8}\right)\right\|,$$

т.е. неравенство $||(a_1,b_1)|| + ||(a_1,b_1) + (a_2,b_2)|| \geqslant ||(a_2,b_2)||$ не выполняется.

Известно, что все конечномерные линейные пространства с несимметричной нормой топологически эквивалентны [33].

Однако класс нормированных конусов значительно шире. Оказывается, что даже в двухмерном линейном пространстве можно определить топологически неэквивалентные нормированные конусы, не допускающие линейного инъективного изометричного вложения ни в какое нормированное пространство. Приведём простые примеры.

Пример 5.7.41. Пусть Q — набор числовых пар $Q = \{(a,b) \,|\, a \geqslant 0, b \geqslant 0 \text{ и } a = 0 \Leftrightarrow b = 0\}$ с нормой $\|(a,b)\|_Q = \max\left\{a^2/b, b^2/a\right\}$ при $(a,b) \neq (0,0)$ и $\|(0,0)\|_Q = 0$.

Рассмотрим естественно возникающую конус-топологию на Q, которая порождена семейством окрестностей $B_\varepsilon(x) = \{x + h \in Q \,|\, h \in Q, \|h\|_Q \leqslant \varepsilon\}$ точек $x \in Q$ при $\varepsilon > 0$. Эта конус-топология не равносильна стандартной топологии в $\mathbb{R}^2$. Действительно, если $x_n = (a_n, b_n) = (1/n, 1)$, то в $\mathbb{R}^2$ верно $\lim\limits_{n\to\infty} x_n = (0,1)$, но $\lim\limits_{n\to\infty} \|x_n\|_Q = +\infty$. Поэтому Q со введённой выше конус-топологией невозможно непрерывно вложить в $\mathbb{R}^2$.

Пример 5.7.42. Аналогичный пример можно рассмотреть на множестве Q числовых пар (a,b), где $a \geqslant 0$, $b \in \mathbb{R}$, причём $b = 0$ означает $a = 0$. Норма в Q вводится следующим образом: $\|(a,b)\|_Q = a + b^2/a$ при $a \neq 0$ и $\|(0,0)\|_Q = 0$. Отметим, что единичный шар $B_Q(0) = \{(a,b) \in Q \mid \|(a,b)\|_Q \leqslant 1\}$ есть круг радиуса $1/2$ с центром в точке $(0, 1/2)$.

Замечание 5.7.43. Все основные результаты данного пункта сформулированы в классе нормированных конусов. Однако нетрудно перенести их на конусы Q с полунормой $\|\cdot\|$, если вместо Q рассматривать множество $Q \setminus Q_0$, где $Q_0 = \{x \in Q \,|\, \|x\| \neq 0\}$ и модифицировать понятие сопряжённого (двойственного) конуса согласно замечанию 5.7.20.

5.7.6 Некоторые условия разрешимости минимизационных задач в нормированных конусах

Напомним, что для использования алгоритмических методов (см. раздел 5.7.2) для задач выпуклого программирования с относительной точностью часто (если $0 \notin Q$) необходимо подобрать начальную точку x^0 так, чтобы $\|x_*\|_K \geqslant \|x^0\|_K$. Например, это возможно при $\|x^0\|_K = \min\{\|x\|_K \,|\, x \in Q\}$. В этой связи важным вопросом является выяснение условий на Q, при которых $\|x\|_K$ достигает минимума Q. Ниже мы приведем несколько таких условий [56], связанных с сублинейным аналогом $*$-слабой топологии в нормированных конусах и соответствующим вариантом теоремы Банаха – Алаоглу.

Недавно в [172] введён аналог сопряжённого пространства к нормированному конусу Q — субсопряжённый конус $\widehat{Q}^*_{sub}$ — набор функционалов $p_\ell : Q \to \mathbb{R}$ вида

$$p_\ell(x) = \max\{0, \ell(x)\}, \tag{5.69}$$

где $\ell : Q \to \mathbb{R}$ — линейны и ограничены по норме ($\ell(x) \leqslant C||x|| \ \forall x \in Q$ при некотором $C > 0$), причем для каждого такого ℓ существует $x_0 \neq 0$, для которого $\ell(x_0) \geqslant 0$.

На множестве $\widehat{Q}^*_{sub}$ можно ввести следующие операции сложения и умножения на неотрицательный скаляр: $p_{\ell_1} \oplus p_{\ell_2} := p_{\ell_1+\ell_2}$, $\lambda \cdot p_\ell := p_{\lambda\cdot\ell}$ для произвольных ℓ_1, ℓ_2 и $\lambda \geqslant 0$. Нетрудно показать, что для введенных операций $\widehat{Q}^*_{sub}$ будет выпуклым конусом с нормой:

$$||p_\ell||_{\widehat{Q}^*_{sub}} := \sup_{x \neq 0} \frac{p_\ell(x)}{||x||}.$$

Введём в конусе $\widehat{Q}^*_{sub}$ $*$-слабую топологию, базу которой образуют окрестности элемента $p_\ell \in \widehat{Q}^*_{sub}$ вида

$$V_{\varepsilon;x_1,x_2,\ldots,x_n}(p_\ell) = \left\{ p_g \in \widehat{Q}^*_{sub} \mid |p_g(x_k) - p_\ell(x_k)| < \varepsilon \ \forall k = \overline{1,n} \right\} \quad (5.70)$$

для фиксированного числа $\varepsilon > 0$ и элементов $x_1, x_2, \ldots, x_n \in Q$. Отметим, что $\widehat{Q}^*_{sub}$ — *хаусдорфово топологическое пространство* в $*$-слабой топологии. Справедлива следующая [56]

Теорема 5.7.44. *Множество* $B^* = B_{\widehat{Q}^*_{sub}} = \left\{ p_\ell \in \widehat{Q}^*_{sub} \mid ||p_\ell||_{\widehat{Q}^*_{sub}} \leqslant 1 \right\}$ *есть компакт в* $*$*-слабой топологии.*

Доказательство. Можно показать, что $p_\ell(x) \in [0; ||x||] \ \forall x \in Q$, $p_\ell \in B^*$. Далее необходимо рассмотреть декартово произведение $H := \prod\limits_{x \in Q} [0; ||x||]$, которое компактно по теореме Тихонова в топологии прямого произведения. Сужение этой топологии на подмножество $T := \bigcup\limits_{p_\ell \in B^*} \{p_\ell(x)\}_{x \in Q} \subset H$ совпадает со $*$-слабой топологией в $\widehat{Q}^*_{sub}$. Замкнутость T можно проверить непосредственно с использованием аналога теоремы Хана–Банаха о продолжении линейного функционала с сохранением выпуклой оценки в классе абстрактных выпуклых конусов ([171], теорема 2.1). □

Выше показано, что всякий отделимый нормированный конус можно метризовать, то есть в нем существует однородная метрика $d_Q : Q \times Q \to \mathbb{R}$ такая, что $d(0, x) = ||x||$ для всякого $x \in Q$. Заметим, что при этом в нормированном пространстве $(E, ||\cdot||)$ для всяких $x, y \in E$ будет верно неравенство $1/2||x - y|| \leqslant d_E(x, y) \leqslant ||x - y||$.

Оказывается, что если $(Q,\ d_Q)$ — сепарабельное метрическое пространство, то $*$-слабую топологию в субсопряженном конусе $\widehat{Q}^*_{sub}$ также можно метризовать и вывести из теоремы 5.7.44 следующее утверждение.

Теорема 5.7.45. *Если (Q, d_Q) — сепарабельное метрическое пространство, то $B^* = \{p_\ell \in \widehat{Q}^*_{sub} \mid ||p_\ell||_{\widehat{Q}^*_{sub}} \leqslant 1\}$ — компакт в некотором метрическом пространстве.*

На базе предыдущего результата можно получить следующее утверждение о существовании минимимума функционала, заданного на $*$-замкнутом подмножестве конуса $Q = \widehat{Y}^*_{sub}$ для некоторого отделимого d_Y-сепарабельного нормированного конуса Y.

Теорема 5.7.46. *Пусть $Q = \widehat{Y}^*_{sub}$, Y — отделимый d_Y-сепарабельный нормированный конус, $f : A \to \mathbb{R}$ — собственный функционал ($f(x) > -\infty\ \forall x \in A$), непрерывный на замкнутом выпуклом множестве A в стандартной конус-топологии, порождённой системой окрестностей $B_\varepsilon(x) = \{x + h \mid \|h\| \leqslant \varepsilon,\ h \in Q\}$ точек $x \in Q$, $\varepsilon > 0$. Тогда существует такой элемент $a_0 \in A$, что $f(a_0) = \min\limits_{x \in A} f(x)$, а также последовательность $\{a_k\}_{k=1}^{\infty} \in A$, для которой $f(a_0) = \lim\limits_{k \to +\infty} f(a_k)$.*

Замечание 5.7.47. Отметим, что функционал f может быть собственным в конусе Q, но несобственным во всяком нормированном пространстве $E_Q \supset Q$. К примеру, этот будет верно для функционала $f((a, b)) = -b^2/a$ в нормированных конусах из примеров 5.7.41 и 5.7.42.

Заключительные замечания к главе 5. Адаптивные зеркальные спуски с использованием δ-субградиентов

В отличие от предыдущих частей работы, настоящая глава посвящена методам зеркального спуска, оптимальность которых не удаётся математически обосновать на классе выпуклых гладких оптимизационных задач. Оптимальность данной методики с точки зрения нижних оценок обоснована для задач выпуклого программирования с липшицевыми (вообще говоря, негладкими) целевыми функционалами и функционалами ограничений. Разработанная методика основана на сочетание

адаптивных подходов к выбору шага и адаптивных критериев остановки методов, которые автоматически гарантируют достижение желаемого качества решения оптимизационной задачи по функции. При этом, в отличие от методов для задач оптимизации без ограничений, пока не удалось предложить адаптивную настройку на величины погрешностей задания функционалов или их (суб)градиентов.

Выделим преимущества предложенных алгоритмических схем:

- оптимальная на классе выпуклых липшицевых функционалов оценка $O(\varepsilon^{-2})$ скорости сходимости зеркального спуска доказана на более широком классе гёльдеровых целевых функционалов, а также для целевых функционалов;

- оценку $O(\varepsilon^{-2})$ скорости сходимости зеркального спуска удаётся перенести и на другие классы целевых функционалов, не удовлетворяющих условию Липшица (например, максимум квадратичных функционалов);

- для некоторых из предложенных в настоящей главе методов доказанные оценки скорости сходимости сохранятся и на классе квазивыпуклых целевых функционалов;

- возможно игнорирование на непродуктивных шагах (на которых нарушено требование к значениям функциональных ограничений) некоторых из функционалов ограничений задачи при сохранении гарантированной оценки качества решения.

Отметим, что во многих задачах операция нахождения точного вектора субградиента целевого функционалам может быть достаточно затратной. Одним из подходов к указанной проблеме может быть использование вместо субградиентов f *δ-субградиентов* в текущей точке x $\nabla_\delta f(x)$:

$$f(y) - f(x) \geqslant \langle \nabla_\delta f(x), y - x \rangle - \delta$$

для всякого $y \in Q$ при некотором фиксированном $\delta \geq 0$ (ясно, что при $\delta = 0$ $\nabla_\delta f(x) = \nabla_\delta f(x)$ — обычный субградиент). Например, для

$$f(x) = \max_{y \in P} \varphi(x, y)$$

выпуклой функции φ по x и непрерывной по $y \in P$ (P — некоторый компакт) $\nabla_\delta f(x)$ есть субградиент по x функционала $\varphi(x, y_\delta)$, где

$$\max_{y \in P} \varphi(x, y) - \varphi(x, y_\delta) \leqslant \delta.$$

В рассматриваемых нами методах зеркального спуска можно заменить обычный субградиента $\nabla_\delta f(x)$ на δ-субградиент $\nabla_\delta f(x)$. Оценки качества решения при этом изменятся на величину $O(\delta)$. Начнём с важного вспомогательного утверждения, которое позволяет получать оценки скорости сходимости для широкого класса целевых функционалов. Пусть при x и $y \in Q$

$$v_f^\delta(x, y) = \left\langle \frac{\nabla_\delta f(x)}{\|\nabla_\delta f(x)\|_*}, x - y \right\rangle$$

при $\nabla_\delta f(x) \neq 0$ и $v_f^\delta(x, y) = 0$ при $\nabla_\delta f(x) = 0$.

Лемма 5.7.48. *Если $f : Q \to \mathbb{R}$ выпуклый функционал и при $\tau > 0$*

$$\omega(\tau) = \max_{x \in Q} \{f(x) - f(x_*) \,|\, \|x - x_*\| \leqslant \tau\}.$$

Тогда для всякого $x \in Q$

$$f(x) - f(x_*) \leqslant \omega(v_f^\delta(x, x_*)) + \delta.$$

Доказательство. 1)Покажем, что

$$v_f^\delta(x, x_*) = \min_y \{\|y - x_*\| \,|\, \langle \nabla_\delta f(x), y - x \rangle = 0\}.$$

Выберем y_* так, чтобы $\|y_* - x_*\| = \min\limits_y \{\|y - x_*\| \,|\, \langle \nabla_\delta f(x), y - x \rangle = 0\}$. Далее, для фиксированного x верно $\langle \nabla_\delta f(x), y_* - x \rangle = 0$. Поэтому существует такое $\lambda > 0$, что $\nabla_\delta f(x) = \lambda s$ при $\langle s, y_* - x_* \rangle = \|y_* - x_*\|$ и $\|s\|_* = 1$. Имеем:

$$0 = \langle \nabla_\delta f(x), y_* - x \rangle = \lambda \langle s, y_* - x_* \rangle + \langle \nabla_\delta f(x), x_* - x \rangle,$$

т.е.

$$\lambda = \frac{\langle \nabla_\delta f(x), x - x_* \rangle}{\|y_* - x_*\|} = \|\nabla_\delta f(x)\|_*.$$

Это и означает, что $v_f^\delta(x, x_*) = \|y_* - x_*\|$.

2) Поскольку $\langle \nabla_\delta f(x), y - x\rangle = 0$, то $f(y) - f(x) \geqslant -\delta$ для всякого y, при котором $\langle \nabla_\delta f(x), y - x\rangle = 0$. Далее,

$$f(x) - f(x_*) \leqslant f(y) - f(x_*) + \delta \leqslant \omega(v_f^\delta(x, x_*)) + \delta,$$

что и требовалось. □

Сформулируем также следующий аналог базовой леммы для зеркальных спусков для δ-субградиентов $\nabla_\delta f$ [**?**].

Лемма 5.7.49. *Пусть f — некоторая выпуклая функция на Q, $h > 0$ — размер шага. Пусть точка y определяется формулой $y = \mathrm{Mirr}[x](h \cdot (\nabla_\delta f(x)))$. Тогда для всякого $z \in Q$*

$$h \cdot \big(f(x) - f(z)\big) \leqslant h \cdot \langle \nabla_\delta f(x), x - z\rangle + h \cdot \delta \leqslant$$

$$\leqslant \frac{h^2}{2}\|\nabla_\delta f(x)\| + h \cdot \delta + V(z, x) - V(z, x^+).$$

Обозначим через $\varepsilon > 0$ фиксированную точность, x_0 — начальное приближение такое, что для некоторого $\Theta_0 > 0$ верно неравенство $V(x_*, x^0) \leqslant \Theta_0^2$. Пусть

$$|g(x) - g(y)| \leqslant M_g||x - y|| \ \forall x, y \in Q.$$

Рассмотрим следующий метод.

Тогда справедлив следующий результат об оценке качества найденного решения предложенного метода.

Теорема 5.7.50. *После остановки предложенного алгоритма 28 справедливы неравенства:*

$$f(\widehat{x}) - f(x_*) \leqslant \varepsilon + \delta \ \textit{и} \ g(\widehat{x}) \leqslant \varepsilon M_g + \delta.$$

Алгоритм 28 Адаптивный зеркальный спуск

Require: $\varepsilon > 0,\ \Theta_0 : d(x_*) \leqslant \Theta_0^2$

$x^0 = argmin_{x \in Q}\, d(x)$

$I =: \emptyset$

$k \leftarrow 0$

repeat

  if $g(x^k) \leqslant \varepsilon ||\nabla_\delta g(x^k)||_* + \delta$ **then**

    $h_k \leftarrow \frac{\varepsilon}{||\nabla_\delta f(x^k)||_*^2}$

    $x^{k+1} \leftarrow Mirr_{x^k}(h_k \nabla_\delta f(x^k))$ // *"продуктивные шаги"*

    $k \to I$

  else

    $h_k \leftarrow \frac{\varepsilon}{||\nabla_\delta g(x^k)||_*}$

    $x^{k+1} \leftarrow Mirr_{x^k}(h_k \nabla_\delta g(x^k))$ // *"непродуктивные шаги"*

  end if

  $k \leftarrow k + 1$

until

$$\frac{2\Theta_0^2}{\varepsilon^2} \leqslant \sum_{k \in I} \frac{1}{||\nabla_\delta f(x^k)||_*^2} + |J|, \tag{5.71}$$

где $|J|$ — количество непродуктивных шагов (мы обозначим через $|I|$ количество продуктивных шагов, то есть $|I| + |J| = N$).

Ensure: $\widehat{x} = \frac{1}{\sum\limits_{k \in I} h_k} \sum\limits_{k \in I} h_k x^k$.

Алгоритм 29 Адаптивный зеркальный спуск

Require: точность $\varepsilon > 0$; начальная точка $x^0; \Theta_0; Q; d(\cdot)$.

$I := \emptyset$
$N \leftarrow 0$
repeat
  if $g(x^N) \leqslant \varepsilon + \delta$ **then**
    $h_N \leftarrow \frac{\varepsilon}{\|\nabla_\delta f(x^N)\|_*}$
    $x^{N+1} \leftarrow Mirr_{x^N}(h_N \nabla_\delta f(x^N))$ ("продуктивные шаги")
    $N \to I$
  else
    $h_N \leftarrow \frac{\varepsilon}{\|\nabla_\delta g(x^N)\|_*^2}$
    $x^{N+1} \leftarrow Mirr_{x^N}(h_N \nabla_\delta g(x^N))$ ("непродуктивные шаги")
  end if
  $N \leftarrow N + 1$
until $\Theta_0^2 \leqslant \frac{\varepsilon^2}{2}\left(|I| + \sum\limits_{k \notin I} \frac{1}{\|\nabla_\delta g(x^k)\|_*^2}\right)$

Ensure: $\bar{x}^N := \arg\min\limits_{x^k, k \in I} f(x^k)$

Теорема 5.7.51. *Пусть известна константа $\Theta_0 > 0$ такая, что $d(x_*) \leqslant \Theta_0^2$. Если $\varepsilon > 0$ — фиксированное число, то алгоритм 29 работает не более*

$$N = \left\lceil \frac{2 \max\{1, M_g^2\} \Theta_0^2}{\varepsilon^2} \right\rceil$$

итераций, причём после его остановки справедливо неравенство

$$\min_{k \in I} v_f(x^k, x_*) \leqslant \varepsilon, \ \max_{k \in I} g(x^k) \leqslant \varepsilon + \delta.$$

Теорема 5.7.52. *Пусть $\varepsilon > 0$ — фиксированное число и выполнен критерий остановки алгоритма 30. Тогда*

$$\min_{k \in I} v_f(x^k, x_*) \leqslant \varepsilon, \ \max_{k \in I} g(x^k) \leqslant \varepsilon + \delta.$$

Следствие 5.7.53. *Предположим, что $f(x) = \max\limits_{i=\overline{1,m}} f_i(x)$, где f_i дифференцируемы в каждой точке $x \in Q$ и*

$$||\nabla_\delta f_i(x) - \nabla_\delta f_i(y)||_* \leqslant L_i ||x - y|| \quad \forall x, y \in Q.$$

Алгоритм 30 Адаптивный зеркальный спуск, постоянное количество итераций.

Require: $\varepsilon > 0, \Theta_0 : d(x_*) \leqslant \Theta_0^2$
$x^0 = argmin_{x \in Q} d(x)$
$I =: \emptyset$
$N \leftarrow 0$
repeat
  if $g(x^N) \leqslant \varepsilon \|\nabla_\delta g(x^N)\|_* + \delta$ **then**
    $M_N = \|\nabla_\delta f(x^N)\|_*$, $h_N = \frac{\varepsilon}{M_N}$
    $x^{N+1} = Mirr_{x^N}(h_N \nabla_\delta f(x^N))$ // *"продуктивные шаги"*
    $N \to I$
  else
    $M_N = \|\nabla_\delta g(x^N)\|_*$, $h_N = \frac{\varepsilon}{M_N}$
    $x^{N+1} = Mirr_{x^N}(h_N \nabla_\delta g(x^N))$ // *"непродуктивные шаги"*
  end if
  $N \leftarrow N + 1$
until $2\frac{\Theta_0^2}{\varepsilon^2} \leqslant N$
Ensure: $\bar{x}^N := argmin_{x^k,\, k \in I} f(x^k)$

Тогда после

$$N = \left\lceil \frac{2 \max\{1, M_g^2\} \Theta_0^2}{\varepsilon^2} \right\rceil$$

шагов работы алгоритма 29 будет верно следующее неравенство:

$$\min_{k \in I} f(x^k) - f(x_*) \leqslant \|\nabla_\delta f(x_*)\| \varepsilon + \frac{L}{2} \varepsilon^2 + \delta,$$

enditemize где $L = \max\limits_{i=\overline{1,m}} L_i$.

Следствие 5.7.54. *Пусть f удовлетворяет условию Липшица*

$$|f(x) - f(y)| \leqslant M_f \|x - y\| \quad \forall x, y \in Q$$

на Q. Тогда после остановки алгоритма 30 будет верно следующее неравенство:

$$\min_{k \in I} f(x^k) - f(x_*) \leqslant M_f \varepsilon + \delta.$$

По-видимому, разработанные в настоящем разделе схемы зеркального спуска можно применять и в случае неточности вспомогательных

операций проектирования согласно концепциям вида (0.11) (в таком случае полученные оценки скорости сходимости изменятся лишь на величины $O(\widetilde{\delta})$).

На базе полученных результатов для зеркальных спусков с переключениями были получены оценки скорости сходимости для задач выпуклой однородной минимизации с относительной точностью при наличии функциональных ограничений-неравенств. Использование именно схем с переключениями может в некоторых случаях позволить избежать дополнительной задачи проектирования 0 на допустимое множество задачи. Также с использованием анализа в абстрактных полунормированных конусах показано, как можно обойти предположение о том, что 0 есть внутренняя точка субдифференциала целевой функции в нуле.

В завершении отметим, что предложены также алгоритмические схемы для относительно липшицевых целевого функционала и ограничения. Как отмечено в разделе 1.5, такая общность существенно расширяет класс задач, для которых возможно использовать полученные оценки скорости сходимости для зеркальных спусков. Однако для таких классов задачи удалось разработать лишь частично адаптивные методы: шаги выбираются постоянными и зависят от констант относительной липшицевости M_f и M_g, но критерий остановки адаптивен. Это дает возможность применять предложенные подходы в случаях, когда нет возможности гарантировать сильную выпуклость дивергенции Брегмана относительно нормы. В частности, это использовано для вывода оценки скорости сходимости для задачи минимизации с относительной точностью однородного выпуклого функционала при более общих предположениях по сравнению с [38]. Рассмотрены также методы с адаптивными продуктивными шагами, но с постоянными непродуктивными. Это даёт возможность применять указанные методы для задач с относительно липшицевыми функционалами ограничений [125, 157].

Основные результаты данной главы опубликованы в работах [51, 57–59, 68, 169, 171, 173, 176, 180].

ЗАКЛЮЧЕНИЕ

Основная часть монографии посвящена изложению новых результатов по адаптивным и универсальным методам для решения задач выпуклого программирования (с целевыми функционалами различного уровня гладкости), выпукло-вогнутых седловых задач, алгоритмическим методам градиентного типа для решения задач выпуклой оптимизации с неточной информацией о целевом функционале и/или его (суб)градиенте. Наиболее важные из представленных результатов ранее опубликованы в рецензируемых изданиях (журналы и сборники трудов международных конференций). Однако настоящее издание представляет из себя более целостное и подробное изложение этих результатов на русском языке, содержит уточнение некоторых немаловажных деталей (которые не было возможности прописать в статьях по причине ограниченности объёма), а также исправление допущенных опечаток в прежних публикациях [52, 53, 55, 179].

Все рассмотренные методы используют информацию первого порядка (значения функций и (суб)градиенты). Интерес к такого типа алгоритмическим процедурам возрос в оптимизации в последние десятилетия ввиду отсутствия больших затрат памяти и независимости оценок вычислительных гарантий от размерности задачи, что важно для возникающих в последнее время задач в конечномерных пространствах больших размерностей. Если заранее не известно точное решение исследуемой оптимизационной задачи, то вполне разумно для оценки качества найденного в ходе итеративного процесса приближённого решения использовать теоретические оценки скорости сходимости. В этой связи численный метод вполне разумно оценивать с точки зрения оптимальность оценки скорости сходимости на выделенном классе оптимизационных задач. Подчеркнём, что в этом плане нас интересовали методы, которые могут гарантировать оценки сложности (количества итераций, которое необходимо выполнить для достижения приемлемого качества решения задачи), которые не зависят от величины размерности пространства рассматриваемых переменных. Как известно, это вполне возможно, если рассматривать (вместо условий на размерность) глобальные характеристики отображений: μ-сильная выпуклость при

$\mu \geqslant 0$, гладкость (липшицевость градиента), гёльдеровость градиента, липшицевость целевого функционала (ограниченность субградиента и т.д. Полученные оценки скорости сходимости предложенных методов сопоставляются с известными нижними оценками на классах задач. Такой подход восходит к известной монографии А.С. Немировского и Д.Б. Юдина [36]. Стоит отметить, что выполненные в настоящей работе исследования нацелены, прежде всего, на разработку эффективных методов для задач с функциональными ограничениями произвольной структуры (т.е. для функционалов ограничений мы в общем случае ограничиваемся лишь информацией о классе рассматриваемых оптимизационных задач) и поэтому для нас важны именно оптимальность оценок именно на выделенных классах задач (а не для конкретной поставленной задачи).

Основной упор в проведённых исследованиях был сделан на разработку методов для негладких оптимизационных задач, выпукло-вогнутых седловых задач (седловая постановка — это также по сути негладкая задача) и вариационных неравенств на базе следующих основных подходов. Первый из них основан на идеологии универсальных методов и заключается в построении для негладкой задачи свойственной для гладкой ситуации оптимизационной модели искусственно введёнными параметрами неточностей с последующей адаптивной настройкой в ходе работы метода на величины этих параметров. Второй подход связан с вариантами абстрактных понятий (концепций) неточной оптимизационной модели для выделенных классов задач. Модельная общность, в частности, позволяет для некоторых классов негладких задач (например, композитная оптимизация) получать оценки скорости сходимости, свойственные для задач с более существенными предположениями гладкости. Третий подход нацелен на применение в задачах выпуклого (и квазивыпуклого) программирования и основан на комбинации адаптивного выбора шага и предложенных адаптивных критериях остановки зеркальных спусков (субградиентных схем) с переключениями. Эти критерии остановки могут позволять говорить, в частности, о применимости разработанных методов вне зависимости от глобальных свойств (условие Липшица или Гёльдера целевого функционала), правда без гарантированных оптимальных оценок скорости сходимости. Мы наблюдали этот эффект, например, в разделе 5.6 для задачи максимизации полезности компьютерной сети в случае логарифмической функции

полезности.

В работе рассмотрены также ситуации неточного задания целевого функционала и (суб)градиента. Преимущество первых двух подходов — возможность адаптивной настройки как на уровень (параметры) гладкости, так и некоторых параметры погрешностей данных (главы 3 и 4). В случае третьего подхода (глава 5) адаптивную настройку на величины возможных погрешности реализовать не удалось. Однако показано, что возможно вместо истинного значения субградиентов использовать неточные δ-субградиенты (см. заключительные замечания к главе 5) и в оценках качества решения не будет накопления параметров, соответствующих параметру неточности δ).

Кратко перечислим основные результаты работы.

- Введены аналоги неточного оракула для вариационных неравенств и седловых задач. Для класса задач, допускающих модель указанного типа, предложен адаптивный вариант проксимального зеркального метода. Получена оценка скорости сходимости этого метода, обоснована его оптимальность на соответствующем классе задач. Показано, что в оценке скорости сходимости не накапливаются величины, соответствующие погрешностями задания оператора вариационного неравенства (функционала седловой задачи), а также неточности решения вспомогательных подзадач на итерациях метода. Обоснована применимость данного метода к популярным для вопросов обработки изображений композитным седловым задачам, а также недавно рассмотренному в [65] примеру задачи о совместном использовании ресурсов.

- Впервые предложен универсальный метод для вариационных неравенств с гёльдеровым монотонным оператором. Получены оценки скорости сходимости этого метода для выпукло-вогнутых седловых задач различного уровня гладкости. Доказана оптимальность предложенных процедур как на классе вариационных неравенств с липшицевым оператором, так и на классе вариационных неравенств с ограниченным оператором.

- Предложен новый подход к описанию неслучайных погрешностей для градиентных методов оптимизации и показана его применимость к широкому классу негладких оптимизационных задач. Особенность предложенного подхода — раздельный учёт в оптими-

зационной модели параметров, соответствующих разными типам погрешностей. В гладком случае это могут быть погрешности задания целевого функционала и градиента. В негладком же случае эти параметры можно использовать для описания в некотором смысле отклонений от гладкости. Предложены неускоренный и ускоренный (на более узком классе задач) метод градиентного типа с адаптивной настройкой в оценке скорости сходимости некоторых из этих параметров и показано, что такая адаптивная настройка может повышать качество найденного решения по сравнению с теоретическими оценками. Доказано, что величины, связанные с постоянными погрешностями, не накапливаются в итоговых оценках для неускоренного метода. Для ускоренного метода обоснована возможность избежать накопления и второго параметра. Обоснована применимость разработанн̆ методики к задачам композитной оптимизации с неточным заданием целевой функции, градиента гладкой части или композитного слагаемого. Однако проблема адаптивной настройки как на уровень гладкости, так и на величины погрешностей полностью не решена, т.к. адаптивная настройка происходит по части параметров неточностей. Тем не менее, в случае искусственных неточностей, связанных с негладкостью задачи, все параметры погрешностей поддаются адаптивной настройке и разработанные процедуры оптимальны на соответствующих классах задач. Обоснована применимость с достаточно эффективными вычислительными гарантиями разработанной методики на некотором классе негладких задач оптимизации. Показана применимость неускоренных процедур для относительно гладких целевых функционалов. В таком случае полученную оценку скорости сходимости $O(\varepsilon^{-1})$ можно считать оптимальной даже при отсутствии погрешностей [92]. Показано, как можно ввести аналогичное понятие неточной модели для вариационных неравенств и седловых задач и предложить аналог экстраградиентного метода с адаптивной настройкой на величины детерминированной аддитивной погрешности задания оператора.

- Введено новое понятие неточной модели целевой с несколькими параметрами, соответствующими свойствам сильной выпуклости, гладкости, а также погрешностям целевого функционала или градиента (предполагается раздельный учёт этих неточностей). Обос-

нована близкая к линейной скорость сходимости предложенных градиентных методов с адаптивным выбором шага для задач минимизации, допускающих существование указанной модели в произвольной запрошенной точке. Разработан адаптивный градиентный метод и обоснована близкая к линейной скорость его сходимости в случае, если вместо сильной выпуклости целевого функционала относительно евклидовой нормы выполняется условие градиентного доминирования Поляка-Лоясиевича. Отметим, что для указанного метода возможна полная адаптация оценки скорости сходимости к погрешности задания градиента на неограниченном допустимом множестве.

- Предложены зеркальные спуски с переключениями для задач выпуклого программирования с новыми адаптивными критериями остановки, выполнение которых гарантирует достижение приемлемого качества решения вне зависимости от уровня гладкости задачи. Обоснована оптимальность методов для (вообще говоря, не удовлетворяющих условию Липшица) целевых функционалов различного уровня гладкости с липшицевым функционалом ограничения. В частности доказано, что на классе задач с гёльдеровыми выпуклыми целевыми функционалами сохранится оценка сложности $O(\varepsilon^{-2})$, оптимальная даже на более узком классе липшицевых выпуклых целевых функционалов. Рассмотрено приложение к задаче оптимизации компьютерной сети. Показано, что некоторые из разработанных методов применимы и для задачи минимизации квазивыпуклого функционала с сохранением всех полученных для выпуклых задач оценок сложности методов для функционалов соответствующего уровня гладкости (гёльдеров целевой функционал, функционал с липшицевым градиентом). Особенно интересными представляются полученные результаты для сильно выпуклых задач, поскольку в этом случае даже для безусловных задач не известны методы, гарантирующие сохранение оптимальных вычислительных гарантий на столь широких классах негладких функционалов, не удовлетворяющих условию Липшица. Показано, что разработанные методы зеркального спуска с переключениями применимы к задачам минимизации с относительной точностью выпуклого однородного функционала с выпуклыми функционалами ограничений. Если эти ограничения возможно записать

в виде подходящей системы неравенств, то это может позволить избежать дополнительной подзадачи проектирования 0 на допустимое множество. Указанные методы гарантируют достижение заданной относительной точности приближённого решения задачи за оптимальное число итераций при существенно более общих предположениях в сравнении с известными аналогичными результатами (0 — не обязательно внутренняя точка субдифференциала целевой функции в нулевой точке).

Помимо теоретических исследований с привлечением (только для написания кодов) соисполнителей по грантам также проводились вычислительные эксперименты со случайно сгенерированными данными на базе программной реализации методов в среде CPython 3.7 (выполненные совместно со Степановым А.Н. имеются в [60]) с целью иллюстрации возможности повышения качества работы предлагаемых методов (по сравнению с оптимальными оценками) за счёт предлагаемых адаптивных методов выбора шагов, а также адаптивных критериев остановки (постановки задач и подбор входных параметров для экспериментов принадлежат автору монографии). Важно, что обоснованные нами теоретические оценки скорости сходимости методов позволяют при проведении вычислительных экспериментов делать выводы об эффективности найденного методом решения без априорного знания того, где расположена искомая оптимальная точка.

Важная часть диссертационной работы заключалась в разработке адаптивных и универсальных методов для выпукло-вогнутых седловых задач. Однако в последние годы появились работы, в которых обоснованы более эффективные оценки сложности на классе гладких сильно выпукло-вогнутых седловых задач путём сведения их к системе оптимизационных подзадач по разным группам переменных с применением к этим подзадачам оптимальных ускоренных методов оптимизации (см., например [3]). Для одной из этих подзадач применяются методы с неточным (δ, L)-градиентом, причём неточность связана с погрешностью решения вспомогательных подзадач. Как ожидается, в случае, если эти оптимизационные задачи достаточно гладкие, такого типа подходы могут оказаться более эффективными по сравнению с методами экстраградиентного типа, которые исследованы в настоящей работе. Однако ситуация резко меняется, если степень гладкости подзадач более низкая и вспомогательные оптимизационные подзадачи негладкие. Это

связано с тем, что ускоренные методы не приводят к улучшению теоретических оценок сложности для негладких задач. В негладком случае оптимальные оценки для седловых задач можно получить сведением к вариационному неравенству и применением методов экстраградиентного типа (в том числе предложенных в настоящей работе). Представляется, что на будущее весьма интересна задача построения адаптивных вариантов ускоренных методов (с разными концепциями неточности для внешней подзадачи) для седловых задач и сравнения их эффективности с методами экстраградиентного типа.

В плане возможных направлений развития исследований данной работы можно выделить также проблему построения концепций неточного оракула (модели) для методов высоких порядков, рандомизированные варианты предложенных алгоритмических методов, а также исследование применимости и оценок скорости сходимости разработанных методов для различных классов бесконечномерных задач.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Агафонов А. Д., Стонякин Ф. С. Градиентные методы для задач оптимизации, допускающие существование неточной сильно выпуклой модели целевой функции. // Труды МФТИ. 2019. Том 11 (44), № 3. С. 4–19.

[2] Алимов А. Р. Теорема Банаха–Мазура для пространств с несимметричным расстоянием. // Успехи мат. наук. 2003. Т. 58, № 2. С. 159–160.

[3] Алкуса М. С., Гасников А. В., Двинских Д.М., Ковалёв Д. А., Стонякин Ф. С. Ускоренные методы для седловых задач. // Журнал вычислит. математ. и мат. физики. — т. 60, 2020. — В печати. Доступно по ссылке `https://arxiv.org/pdf/1906.03620.pdf`.

[4] Антипин А. С. Методы решения вариационных неравенств со связанными ограничениями. // Журнал вычислит. матем. и матем. физ. 2000. Т. 40, № 9. С. 1291–1307.

[5] Антипин А. С. Равновесное программирование: методы градиентного типа. // Автомат. и телемех. 1997. Т. 58, № 8. С. 125–137.

[6] Антипин А. С. Равновесное программирование: проксимальные методы // Журнал вычислит. матем. и матем. физ. 1997. Т. 37, № 11. С. 1327–1339.

[7] Антипин А. С., Васильев Ф. П. Регуляризованный метод с прогнозом для решения вариационных неравенств с неточно заданным множеством. // Журнал вычислит. матем. и матем. физ. 2004. Т. 44, № 5. С. 796–804.

[8] Антипин А. С., Васильев Ф. П. Методы регуляризации для решения неустойчивых задач равновесного программирования со связанными ограничениями. // Журнал вычислит. матем. и матем. физ. 2005. Т. 45, № 1. С. 27–40.

[9] Антипин А. С., Васильев Ф. П., Шпирко С. В. Регуляризованный экстраградиентный метод решения задач равновесного программирования с неточно заданным множеством. // Журнал вычислит. матем. и матем. физ. 2005. Т. 45, № 4. С. 650–660.

[10] Антипин А. С., Ячимович В., Ячимович М. Динамика и вариационные неравенства. // Журнал вычисл. матем. и матем. физ. 2017. Т. 57, № 5. С. 783–800.

[11] Баяндина А. С., Гасников А. В., Гасникова Е. В., Мациевский С. В. Прямо-двойственный метод зеркального спуска для условных задач стохастической оптимизации. // Журнал вычисл. матем. и матем. физ. 2018. Т. 58, № 11. С. 1794–1803.

[12] Бородин П. А. Теорема Банаха–Мазура для пространств с несимметричной нормой и ее приложения в выпуклом анализе. Матем. заметки. 2001. Т. 69, № 3. С. 329–337.

[13] Васильев Ф. П. Методы оптимизации. Кн. 1. М.: МЦНМО. 2011. — 624 с.

[14] Васильев Ф. П. Методы оптимизации. Кн. 2. М.: МЦНМО. 2011. — 434 с.

[15] Ведель Я. И., Семенов В. В. Новый двухэтапный проксимальный алгоритм для решения задачи о равновесии. // Журнал вычислит. и прикладной математики. 2015. № 1(118). С. 15–23.

[16] Гасников А. В. Современные численные методы оптимизации. Метод универсального градиентного спуска. — М. Изд-во МФТИ: 2018. — 160 с. доступно по ссылке: `https://arxiv.org/abs/1711.00394`.

[17] Гасников А. В., Двуреченский П. Е., Стонякин Ф. С., Титов А. А. Адаптивный проксимальный метод для вариационных неравенств. // Журнал вычисл. матем. и матем. физ. 2019. Т. 59, № 5. С. 889–894.

[18] Гасников А. В., Камзолов Д. И., Мендель М. А. Основные конструкции над алгоритмами выпуклой оптимизации и их приложения к получению новыхоценок для сильно выпуклых задач. ТРУДЫ МФТИ. 2016. Том 8, №3. С. 25–42.

[19] Гасников А. В., Ковалёв Д. А., Мохаммед А. А. М., Черноусова Е. О. Обоснование гипотезы об оптимальных оценках скорости сходимости численных методов выпуклой оптимизации высоких порядков. // Компьютерные исследования и моделирование. 2018, № 6. — С. 737–753.

[20] Гасников А. В., Крымова Е. А., Лагуновская А. А., Усманова И. Н., Федоренко Ф. А. Стохастическая онлайн оптимизация. Одноточечные и двухточечные нелинейные многорукие бандиты. Выпуклый и сильно выпуклый случаи. // Автомат. и телемех. 2017. № 2. С. 36–49.

[21] Гасников А. В., Лагуновская А. А., Морозова Л. Э. О связи имитационной логит-динамики в популяционной теории игр и метода зеркального спуска в онлайн оптимизации на примере задачи выбора кратчайшего маршрута. // Труды МФТИ. 2015. Т. 7. № 4. С. 104–113.

[22] Гасников А. В., Тюрин А. И. Быстрый градиентный спуск для задач выпуклой минимизации с оракулом, выдающим (δ, L)-модель функции в запрошенной точке. // Журнал вычисл. матем. и матем. физ. 2019. Т.59, № 7. С. 1137–1150.

[23] Данскин Дж. М. Теория максимина и ее приложение к задачам распределения вооружения. — М.: Изд-во «Советское радио». 1970. — 200 с.

[24] Демьянов В. Ф., Малоземов В. Н. Введение в минимакс. — М.: Наука. 1972. — 368 с.

[25] Долженко Е. П., Савостьянов Е. А. Аппроксимации со знакочувствительным весом. Изв. РАН. Сер. матем. 1999. Т. 63, № 6. С. 77–118.

[26] Иванова А. С., Пасечнюк Д. А., Двуреченский П. Е., Гасников А. В., Воронцова Е. А. Численные методы для задачи распределения ресурсов в компьютерной сети. // arXiv:1909.13321v2 (2019). Журнал высыслит. математ. и мат. физики. — Принято в печать (2021 год, № 2).

[27] Иванов Г. Е., Лопушански М. С. О корректности задач аппроксимации и оптимизации для слабо выпуклых множеств и функций. Фундаментальная и прикладная математика. 2013. Т.18, № 5. С. 89–118.

[28] Измаилов А. Ф., Третьяков А. А. Фактор-анализ нелинейных отображений. — М.: Наука. 1994. — 336 с.

[29] Иоффе А. Д., Тихомиров В. М. Теория экстремальных задач. — М.: Наука. 1974. — 460 с.

[30] Кларк Ф. Оптимизация и негладкий анализ: Пер. с англ. — М.: Наука. 1988. — 279 с.

[31] Коннов И. В., Салахутдин Р. А. Двухуровневый итеративный метод для нестационарных смешанных вариационных неравенств. // Изв. вузов. Матем. 2017. № 10. С. 50-61.

[32] Корпелевич Г. М. Экстраградиентный метод для отыскания седловых точек и других задач. // Экономика и матем. методы. 1976. Т. 12. № 4. С. 747–756.

[33] Крейн М. Г., Нудельман А. А. Проблема моментов Маркова и экстремальные задачи. — М.: Наука. 1973. — 551 с.

[34] Люстерник Л. А.,Соболев В. И. Краткий курс функционального анализа. — М.: Высш. школа. 1982. — 271 с.

[35] Меленьчук Н. В. Методы и алгоритмы для решения задач математического моделирования на основе вариационных неравенств. // Дисс. ...канд. физ.-мат. наук: 05.13.18. — Омск: 2011. — 123 с.

[36] Немировский А. С., Юдин Д. Б. Сложность задач и эффективность методов оптимизации. — М.: Наука. Главная редакция физико-математической литературы. 1979. — 384 с.

[37] Немировский А. С., Юдин Д. Б. Эффективные методы решения задач выпуклого программирования большой размерности. // Экономика и математические методы. 1979. № 2. С. 135–152.

[38] Нестеров Ю. Е. Алгоритмическая выпуклая оптимизация. // Дисс... д.ф.-м.н.: 01.01.07. — М.: МФТИ. 2013. — 367 с.

[39] Нестеров Ю. Е. Введение в выпуклую оптимизацию. — М. МЦНМО. 2010. — 280 с.

[40] Нестеров Ю. Е. Метод минимизации выпуклых функций со скоростью сходимости $O(1/k^2)$. // Доклады АН СССР. 1983. Т. 269. № 3. С. 543–547.

[41] Нестеров Ю. Е. Методы минимизации для негладких выпуклых и квазивыпуклых функционалов. // Экономика и математ. методы. 1984. Т. 29. С. 519–531.

[42] Нестеров Ю. Е. Эффективные методы в нелинейном программировании. — М.: Радио и связь. 1989. — 301 с.

[43] Орлов И. В. Теоремы об обратной и неявной функциях в классе субгладких отображений. // Матем. заметки. 2016. Т. 99, № 4. С. 631–634.

[44] Пасечнюк Д. А., Стонякин Ф. С. Об одном методе минимизации выпуклой липшицевой функции двух переменных на квадрате, Компьютерные исследования и моделирование. 2019. Т. 11, № 3. С. 379–395.

[45] Поляк Б. Т. Градиентные методы минимизации функционалов. // Журнал вычислительной математики и математической физики. 1963. Т. 3, № 4. С. 643–653.

[46] Поляк Б. Т. Введение в оптимизацию. — М.: Наука. 1983. — 433 с.

[47] Поляк Б. Т. Минимизация негладких функционалов. // Журнал вычисл. матем. и матем. физ. 1969. 9:3. С. 509-521.

[48] Поляк Б. Т. Один общий метод решения экстремальных задач. // Докл. АН СССР. 1967. 174:1. С. 33-36.

[49] Рокафеллар Р. Т. Выпуклый анализ. — М.: Мир. 1973. — 420 с.

[50] Рохлин Д. Б. Распределение ресурсов в сетях связи с большим числом пользователей: стохастический метод градиентного спуска. // Теория вероятностей и её применения. 2020 (в печати).

[51] Стонякин Ф.С., Баран И.В. О некоторых алгоритмах для условных задач оптимизации с относительной точностью по целевому функционалу. // Труды ИММ УрО РАН. 2020. Т. 26, № 3. С. 198–210.

[52] Стонякин Ф. С. Адаптация к величинам погрешностей для некоторых методов оптимизации градиентного типа. // Труды ИММ УрО РАН. 2019. Т. 25, № 4. С. 210–225.

[53] Стонякин Ф. С. Адаптивные градиентные методы для некоторых классов задач негладкой оптимизации // Труды МФТИ. 2020. Т. 12(45), № 1. С. 112–136. // arXiv:1911.08425 (2020). https://arxiv.org/abs/1911.08425v14.

[54] Стонякин Ф. С. Адаптивный аналог метода Ю. Е. Нестерова для вариационных неравенств с сильно монотонным оператором. // Сиб. журн. вычислит. матем. 2019. Т. 22, № 2. С. 201–211.

[55] Стонякин Ф. С. Аналог квадратичной интерполяции для специального класса негладких функционалов и одно его приложение к адаптивному методу зеркального спуска. // Динамические системы. 2019. Т. 9(37), № 1. С. 3–16. ArXiv preprint: https://arxiv.org/abs/1812.04517v2.

[56] Стонякин Ф. С. О сублинейных аналогах слабых топологий в нормированных конусах. // Матем. заметки, т. 103, № 5 (2018), 794–800.

[57] Стонякин Ф. С. Сублинейный аналог теоремы Банаха–Мазура в отделимых выпуклых конусах с нормой. // Матем. заметки. 2018. Т. 104, № 1. С. 118–130.

[58] Стонякин Ф. С., Алкуса М. С., Степанов А. Н., Баринов М. А. Адаптивные алгоритмы зеркального спуска в задачах выпуклого программирования с липшицевыми ограничениями. // Тр. ИММ УрО РАН. 2018. 24, № 2. С. 266—279.

[59] Стонякин Ф. С., Алкуса М., Степанов А. Н., Титов А. А. Адаптивные алгоритмы зеркального спуска для задач выпуклой и сильно выпуклой оптимизации с функциональными ограничениями. // Дискретный анализ и исслед. опер. 2019. Т. 26, № 3. С. 88–114.

[60] Стонякин Ф. С., Степанов А. Н. Исходные коды некоторых вычислительных экспериментов. https://github.com/stonyakin/monograph.

[61] Хачиян Л. Г. Полиномиальный алгоритм в линейном программировании. // Докл. АН СССР. 1979. Т. 244, № 5. С. 1093–1096.

[62] Шор Н. З. Применение обобщённого градиентного спуска в блочном программировании. // Кибернетика. 1967. № 3. С. 53–55.

[63] Allen-Zhu, Z., Hazan, E. Optimal black-box reductions between optimization objectives. // Advances in Neural Information Processing Systems. 2016. P. 1614–1622.

[64] 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:1508.07607 (2015).

[65] Antonakopoulos K., Belmega V., Mertikopoulos P. An adaptive Mirror-Prox method for variational inequalities with singular operators. // Advances in Neural Information Processing Systems. 2019. Vol. 32. P. 8453–8463. https://papers.nips.cc/paper/9053-an-adaptive-mirror-prox-method-for-variational-inequalities-with-singular-operators.pdf.

[66] Bach F., Levy K. A universal algorithm for variational inequalities adaptive to smoothness and noise. // COLT'19: Proceedings of the 32nd Annual Conference on Learning Theory (2019).

[67] Bauschke H., Bolte J., Teboulle M. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. // Mathematics of Operations Research. 2017. Vol. 42, No. 2, P. 330–348.

[68] Bayandina A., Dvurechensky P., Gasnikov A., Stonyakin F., Titov A. Mirror Descent and Convex Optimization Problems With Non-Smooth Inequality Constraints. // Lecture Notes in Mathematics. 2018. Vol. 2227. P. 181–213.

[69] Bao T., Khanh P. Some Algorithms for Solving Mixed Variational Inequalities. // Acta Mathematica Vietnamica. 2006. Vol. 31, No. 1. P. 77–98.

[70] Beck A. First-Order Methods in Optimization MOS-SIAM Series on Optimization. SIAM. 2017. — 467 p.

[71] Beck A., Ben-Tal A., Guttmann-Beck N., Tetruashvili L. The comirror algorithm for solving nonsmooth constrained convex problems. // Operations Research Letters. 2010. Vol. 38, No. 6. P. 493–498.

[72] Beck A., Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. // SIAM J. Imaging Sci. 2009. Vol. 2, No. 1. P. 183–202.

[73] Beck A., Teboulle M. Mirror descent and nonlinear projected subgradient methods for convex optimization. // Oper. Res. Lett. 2003. Vol. 31, No. 3. P. 167–175.

[74] Ben-Tal A., Nemirovski A. Lectures on Modern Convex Optimization. // Philadelphia: SIAM (2001).

[75] Ben-Tal A., Nemirovski A. Lectures on Modern Convex Optimization (Lecture Notes). // Personal web-page of A. Nemirovski (2015).

[76] Ben-Tal A., Nemirovski A. Robust truss topology design via semidefnite programming. // SIAM J. Optim. 1997. Vol. 7, No. 4. P. 991–1016.

[77] Blum L., Cucker F., Shub M., Smale S. Complexity and real computation. // Springer Science & Business Media. 2012.

[78] Boyd S., Vandenberghe L. Convex Optimization. // New York, NY: Cambridge University Press. 2004.

[79] Bogolubsky L., Dvurechensky P., Gasnikov A., Gusev G., Nesterov Y., Raigorodskii A., Tikhonov A., Zhukovskii M. Learning supervised pagerank with gradient-based and gradient-free optimization methods. // Advances in Neural Information Processing Systems 29. 2016. P. 4914–4922.

[80] Brent R. Algorithms for Minimization Without Derivatives. // Dover Books on Mathematics. Dover Publications (1973).

[81] Bubeck S., Eldan R. Multi-scale exploration of convex functions and bandit convex optimization. // e-print (2015).

[82] Bubeck S., Cesa-Bianchi N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. // Foundation and Trends in Machine Learning. 2012. Vol. 5, No. 1. P. 1–122.

[83] Bubeck S. Convex optimization: algorithms and complexity. // Foundations and Trends in Machine Learning. 2015. Vol. 8, issue 3–4. P. 231–357.

[84] Chambolle A., Pock T. A first-order primal-dual algorithm for convex problems with applications to imaging. // Journal of Mathematical Imaging and Vision. 2011. Vol. 40, No. 1. P. 120–145.

[85] Chen Y., Lan G., Ouyang Y. Optimal primal-dual methods for a class of saddle point problems. // SIAM Journal on Optimization. 2014. Vol. 24, No. 4. P. 1779–1814.

[86] Cobzas S. Functional Analysis in Asymmetric Normed Spaces. // Basel. Birkhauser/Springer (2013).

[87] Cohen M., Lee Y., Miller G., Pachocki J., Sidford A. Geometric median in nearly linear time. // Proceedings of the forty-eighth annual ACM symposium on Theory of Computing. 2016. P. 9–21.

[88] Cohen M., Diakonikolas J., Orecchia L. On Acceleration with Noise-Corrupted Gradients. // Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, Proceedings of Machine Learning Research 80 (2018).

[89] D'Aspremont A. Smooth optimization with approximate gradient. // SIAM Journal of Optimization. 2008. Vol. 19, No. 3. P. 1171–1183.

[90] Devolder O., Glineur F., Nesterov Y. First-order methods of smooth convex optimization with inexact oracle. // Math. Program. 2014. Vol. 146, No. 1–2. P. 37–75.

[91] Devolder O. Exactness, Inexactness and Stochasticity in First-Order Methods for Large-Scale Convex Optimization. // PhD thesis (2013).

[92] Dragomir R.-A., Taylor A., dÁspremont A., Bolte J. Optimal Complexity and Certification of Bregman First-Order Methods. // arXiv:1911.08510 (2019).

[93] Drori Y., Teboulle M. An optimal variants of kelley's cutting-plane method. // Mathematical Programming. 2016. Vol. 160, No. 1–2. P. 321–351.

[94] Duchi J. Introductory lectures on stochastic optimization. // Park City Mathematics Institute, Graduate Summer School Lectures (2016).

[95] Duchi J., Bartlett P., Wainwright M. Randomized smoothing for stochastic optimization. // SIAM Journal on Optimization. 2012. Vol. 22, No. 2. P. 674–701.

[96] Dvurechensky P., Gasnikov A. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. // Journal of Optimization Theory and Applications. 2016. Vol. 171, No. 1. P. 121–145.

[97] Dvurechensky P., Dvinskikh D., Gasnikov A., Uribe C. A., Nedic A. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. // Proceedings of the 32th Conference on Neural Information Processing Systems, NIPS'18 (2018).

[98] Dvurechensky P. Gradient method with inexact oracle for composite non-convex optimization. // arXiv:1703.09180 (2017).

[99] Facchinei F., Pang J. S. Finite-Dimensional Variational Inequality and Complementarity Problems. // Springer-Verlag, New York. 2003. Vols. 1 and 2.

[100] Garcia-Raffi L. M., Romaguera S., Sanchez-Perez E. A., Valero O. Metrizability of the unit ball of the dual of a quasi-normed cone. // Bollettino dell'Unione Matematica Italiana. 2004. Vol. 7-B, No. 8. P. 483–492.

[101] Gasnikov A. V., Kabanikhin S. I., Mohamed A., Shishlenin M. A. Convex optimization in Hilbert space with applications to inverse problems. // arXiv:1703.00267 (2017).

[102] Gorbunov E., Dvinskikh D., Gasnikov A. Optimal Decentralized Distributed Algorithms for Stochastic Convex Optimization. // arXiv:1911.07363 (2019).

[103] Guzman C., Nemirovski A. On lower complexity bounds for large-scale smooth convex optimization. // Journal of Complexity. 2015. Vol. 31. P. 1–14.

[104] Hanzely F., Richtarik P. Randomized methods for minimizing relatively smooth functions. // Tech. report (2017).

[105] Hazan E., Kale S. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. // JMLR. 2014. Vol. 15. P. 2489–2512.

[106] Hazan E. Introduction to online convex optimization. // Foundations and Trends in Optimization. 2015. Vol. 2, No. 3–4. P. 157–325.

[107] Hendrikx, Hadrien & Xiao, Lin & Bubeck, Sebastien & Bach, Francis & Massoulie, Laurent. (2020). Statistically Preconditioned Accelerated Gradient Method for Distributed Optimization.

[108] Ivanov G. On well posed best approximation problems for a nonsymmetric seminorm. // Journal of Convex Analysis. 2013. Vol. 20, No. 2. P. 501–529.

[109] Ivanova A., Stonyakin F., Pasechnyuk D, Vorontsova E., Gasnikov A. Adaptive Mirror Descent for the Network Utility Maximization Problem. // IFAC Conference (2020). `https://arxiv.org/abs/1911.07354v1`.

[110] Jenatton R., Huang J., Archambeau C. Adaptive Algorithms for Online Convex Optimization with Long-term Constraints. // arXiv:1512.07422 (2015).

[111] Juditsky A., Nemirovski A. First Order Methods for Non-smooth Convex Large-scale Optimization, I: General purpose methods. // Optimization for Machine Learning, S. Sra et al, Eds., Cambridge, MA: MIT Press. 2012. P. 121–184.

[112] Juditsky A., Nemirovski A., et al. First order methods for non-smooth convex large-scale optimization, ii: utilizing problems structure. // Optimization for Machine Learning. 2011. P. 149–183.

[113] Kalai A., Vempala S. Efficient algorithms for online decision problems. // Journal of Computer and System Sciences. 2005. Vol. 71. P. 291–307.

[114] Karimi H., Nutini J., Schmidt M. Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Lojasiewicz Condition. // Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science. Springer, Cham. 2016. Vol. 9851. P. 795–811.

[115] Karmakar N. A new polynomial time algorithm for linear propramming. // Combinatorica. 1984 Vol. 4, No. 4. P. 373–395.

[116] Keimel K., Roth W. Ordered cones and approximation. // Lecture Notes in Math. Berlin. Springer (1992).

[117] Kelly F., Maulloo A., Tan D. Rate control for communication networks: shadow prices, proportional fairness and stability. // J. Oper Res Soc. 1998. Vol. 49, No. 3. P. 237–252.

[118] Lacoste-Julien S., Schmidt M., Bach F. A simpler approach to obtaining $o(1/t)$ convergence rate for the projected stochastic subgradient method. // arXiv:1212.2002 (2012).

[119] Lagae S. New efficient techniques to solve sparse structured linear systems, with applications to truss topology optimization. Ecole polytechnique de Louvain, Université catholique de Louvain (2017).

[120] Lan G. Gradient sliding for composite optimization. // Mathematical Programming. 2016. Vol. 159, No. 1. P. 201–235.

[121] Lan G., Ouyang Y. Accelerated gradient sliding for structured convex optimization. // arXiv:1609.04905 (2016).

[122] Lee Y., Sidford A., Wong S. C.w. A faster cutting plane method and its implications for combinatorial and convex optimization. // Foundations of Computer Science (FOCS), IEEE 56th Annual Symposium on. P. 1049–1065. IEEE (2015).

[123] Levin A. On an algorithm for the minimization of convex functions. Soviet Math. Doklady (1965).

[124] Lu H., Freund R., Nesterov Y. Relatively smooth convex optimization by Firstorder methods, and applications. // SIAM Journal on Optimization. 2018. Vol. 28, No. 1. P. 333–354.

[125] Lu H. "Relative-Continuity"for Non-Lipschitz Non-Smooth Convex Optimization using Stochastic (or Deterministic) Mirror Descent. // arXiv:1710.04718 (2018).

[126] Lugosi G., Cesa-Bianchi N. Prediction, learning and games. New York, Cambridge University Press (2006).

[127] Mastroeni G. On auxiliary principle for equilibrium problems. // Publicatione del Departimento di Mathematica Dell'Universita di Pisa. 2000. Vol. 3. P. 1244–1258.

[128] Meng X., Chen H. Accelerating Nesterov's Method for Strongly Convex Functions with Lipschitz Gradient. // arXiv:1109.6058 (2011). https://arxiv.org/pdf/1109.6058.pdf.

[129] Monteiro R., Svaiter B. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods // SIAM Journal on Optimization. 2013. Vol. 23, No. 2. P. 1092–1125.

[130] Mordukhovich B., Nam N. Applications of variational analysis to a generalized Fermat-Torricelli problem. // J. Optim. Theory Appl. 2011. Vol. 148, No. 3. P. 431–454.

[131] Mordukhovich B. Variational Analysis And Generalized Differentiation I, Theory And Examples, Grundlehren der mathematischen Wissenschaften, No. 330, Springer (2005).

[132] Necoara I., Nesterov Y. Glineur F. Linear convergence of first order methods for non-strongly convex optimization. // Math. Program. 2019. Vol. 175. P. 69–107.

[133] Nedic A., Ozdaglar A. Subgradient Methods in Network Resource Allocation: Rate Analysis. // Proc. of CISS (2008).

[134] Nemirovsky A. Information-based complexity of linear operator equations. // Journal of Complexity. 1992. Vol. 8, No. 2. P. 153–175.

[135] Nemirovski A. Prox-method with rate of convergence $O(1/T)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. // SIAM Journal on Optimization. 2004. Vol. 15. P. 229–251.

[136] Nemirovski A. Information-based complexity of convex programming. Technion, Fall Semester 1994/95. URL: http://www2.isye.gatech.edu/~nemirovs/Lec_EMCO.pdf.

[137] Nemirovski, A., Onn, S., and Rothblum, U. G. Accuracy certificates for computational problems with convex structure. // Mathematics of Operations Research. 2010. Vol. 35, No. 1. P. 52–78.

[138] Nemirovski A. Lectures on modern convex optimization analysis, algorithms, and engineering applications. Philadelphia: SIAM (2013).

[139] Nesterov Y. Introductory Lectures on Convex Optimization: a basic course. Kluwer Academic Publishers, Massachusetts (2004).

[140] Nesterov Yu. Lectures in Convex Optimization. Springer International Publishing. 2018.

[141] Nesterov Y. Subgradient methods for convex functions with nonstandard growth properties (2016). URL: www.mathnet.ru:8080/PresentFiles/16179/growthbm_nesterov.pdf.

[142] Nesterov Y. Dual extrapolation and its application for solving variational inequalities and related problems. // Math. Program. 2007. Ser. B. P. 319–344.

[143] Nesterov Y., Scrimali L. Solving strongly monotone variational and quasi-variational inequalities. // Discrete and Continuous Dynamical Systems A. 2011. Vol. 31, No 4. P. 1383–1396.

[144] Nesterov Y. Soft clustering by convex electoral model. Core Discussion Paper (2018). Available at https://ideas.repec.org/p/cor/louvco/2018001.html.

[145] Nesterov Y., Polyak B. Cubic regularization of Newton method and its global performance. // Math. Program. Ser. A. 2006. Vol. 108. P. 177–205.

[146] Nesterov Y. Gradient methods for minimizing composite functions. // Math. Program. 2013. Vol. 140, No. 1. P. 125–161.

[147] Nesterov Y., Nemirovskii A. Interior point polynomial methods in convex programming: Theory and Applications. Philadelphia: SIAM, 1994 — 520 p.

[148] Nesterov Y. Primal-dual subgradient methods for convex problems. // Math. Program. 2009. Vol. 120, No. 1. P. 221–259.

[149] Nesterov Y. Smooth minimization of non-smooth functions. // Math. Program. 2005. Vol. 103. P. 127–152.

[150] Nesterov Y. Rounding of convex sets and efficient gradient methods for linear programming problems. // Optimization Methods and Software. Vol. 23, No. 1. P. 109–128.

[151] Nesterov Y. Unconstrained Convex Minimization in Relative Scale. // Mathematics of Operations Research. 2009. Vol. 34, No. 1. P. 180–193.

[152] Nesterov Y. Universal gradient methods for convex optimization problems. // Math. Prog. 2015. Vol. 152, No. 1-2. P. 381–404.

[153] Nesterov Y. Subgradient methods for Huge-Scale Optimization Problems. // Math. Prog. 2015. Vol. 146, No. 1–2. P. 275–297.

[154] Nesterov Yu. Implementable tensor methods in unconstrained convex optimization. // CORE Discussion Papers. 2018/5. — 2018. — 22 p.

[155] Nesterov Yu. Inexact accelerated high-order proximal-point methods. // CORE Discussion Papers. 2020/8. — 2020. — 21 p.

[156] Nesterov Y. Inexact high-order proximal-point methods with auxiliary search procedure. // CORE Discussion paper. — 2020/10. — 23 p.

[157] Nesterov Y. Relative Smoothness: New Paradigm in Convex Optimization. // Conference report, EUSIPCO-2019, A Coruna, Spain, September 4, 2019.

[158] Newman D. Location of the maximum on unimodal surfaces. // Journal of the Association for Computing Machinery. 1965. Vol. 12. P. 395–398.

[159] Nguyen Q. Forward-backward splitting with Bregman distances. // Vietnam Journal of Mathematics. 2017. Vol. 45, No. 3. P. 519–539.

[160] Polyak B., Tremba A. New versions of Newton method: step-size choice, convergence domain and under-determined equations. // Optimization Methods and Software. 2018. P. 1-32. URL: https://arxiv.org/pdf/1703.07810.pdf.

[161] Rädström J. H. An embedding theorem for space of convex sets. // Proc. Amer. Math. Soc. 1952. Vol. 3. P. 165–169.

[162] Romaguera S., Sanchez-Perez E., Valero O. Characterization of Generalized Monotone Normed Cones. // Acta Mathematica Sinica, English Series. 2006. Vol. 23, No. 6. P. 1067–1074.

[163] Roth W. Hahn-Banach type theorems for locally convex cones. // Journal of the Australian Math. Soc. 2000. Vol. 68, No. 1. P. 104–125.

[164] Scaman K., Bach F., Bubeck S., Lee Y., Massoulie L. Optimal Algorithms for Non-Smooth Distributed Optimization in Networks. // Advances in Neural Information Processing Systems 31 (NIPS 2018). URL: https://papers.nips.cc/paper/7539-optimal-algorithms-for-non-smooth-distributed-optimization-in-networks.pdf.

[165] Selinger P. Towards a semantics for higher-order quantum computation. // Proceedings of the 2nd International Workshop on Quantum Programming Languages. Turku Centre for Computer Science General Publication. 2014. Vol. 33. P. 127–143.

[166] Shor N. Minimization Methods for Non-Differentiable Functions. Springer-Verlag Berlin Heidelberg (1985).

[167] Shpirko S., Nesterov Y. Primal-dual Subgradient Methods for Huge-scale Linear Conic Problem. // SIAM Journal on Optimization. 2014. Vol. 24, No. 3. P. 1444–1457.

[168] Stonyakin F., Dvinskikh D., Dvurechensky P., Kroshnin A., Kuznetsova O., Agafonov A., Gasnikov A., Tyurin A., Uribe C. A., Pasechnyuk D., Artamonov S. Gradient Methods for Problems with Inexact Model of the Objective. // In: Khachay M., Kochetov Y., Pardalos P. (eds) Mathematical Optimization Theory and Operations Research. MOTOR 2019. Lecture Notes in Computer Science. Springer, Cham. 2019. Vol. 11548. P. 97–114.

[169] Stonyakin F., Titov A. One Mirror Descent Algorithm for Convex Constrained Optimization Problems with Non-Standard Growth Properties. In Proceedings of the School-Seminar on Optimization Problems and their Applications (OPTA-SCL 2018) Omsk, Russia, July 8-14, 2018. CEUR Workshop Proceedings. 2018. Vol. 2098, P. 372–384.

[170] Stonyakin F. Applications of anticompact sets to analogs of Denjoy-Young-Saks and Lebesgue theorems. // Eurasian. Math. J. 2015. Vol. 6, No. 1. P. 115–122.

[171] Stonyakin F. An analogue of the Hahn-Banach Theorem for functionals on abstract convex cone. // Eurasian. Math. J. 2016. Vol. 7, No. 3, P. 89–99.

[172] Stonyakin F. S. Subdifferential calculus in abstract convex cones. // Constructive Nonsmooth Analysis and Related Topics (dedicated to the memory of V.F. Demyanov). CNSA-2017 Proceedings. 2017. P. 316–319.

[173] Stonyakin F. Hahn-Banach type theorems on functional separation for convex ordered normed cones. // Eurasian Math. J. 2019. Vol. 10, No. 1. P. 59–79.

[174] Stonyakin F., Alkousa M., Titov A., Piskunova V. On Some Methods for Strongly Convex Optimization Problems with One Functional Constraint. // In: Khachay M., Kochetov Y., Pardalos P. (eds) Mathematical Optimization Theory and Operations Research.

MOTOR 2019. Springer, Cham. Lecture Notes in Computer Science. 2019. Vol. 11548. P. 82–96.

[175] Stonyakin F., Vorontsova E., Alkousa M. New Version of Mirror Prox for Variational Inequalities with Adaptation to Inexactness. 10th International Conference on Optimization and Applications, OPTIMA-2019. Communications in Computer and Information Sciences. 2020. Vol. 1145. P. 427–442.

[176] Stonyakin F., Stepanov A., Gasnikov A., Titov A. Mirror Descent for Constrained Optimization Problems with Large Subgradient Values. // Computer Research and Modeling. 2020. Vol. 12, No. 2. P. 301–317.

[177] Stonyakin F., Gasnikov A., Dvurechensky P., Alkousa M., Titov A. Generalized Mirror Prox for Monotone Variational Inequalities: Universality and Inexact Oracle. // arXiv:1806.05140 (2019). https://arxiv.org/abs/1806.05140.

[178] Stonyakin, Fedor & Tyurin, Alexander & Gasnikov, Alexander & Dvurechensky, Pavel & Agafonov, Artem & Dvinskikh, Darina & Pasechnyuk, Dmitry & Artamonov, Sergei & Piskunova, Victorya. Inexact Relative Smoothness and Strong Convexity for Optimization and Variational Inequalities by Inexact Model. // arXiv:2001.09013v3 (2020).

[179] Stonyakin F. S. On the Adaptive Proximal Method for a Class of Variational Inequalities and Related Problems. // Proceedings of the Steklov Institute of Mathematics. 2020. Vol. 309, No. 1. P. S139–S150.

[180] Titov A., Stonyakin F., Gasnikov A., Alkousa M. Mirror Descent and Constrained Online Optimization Problem. // Y. Evtushenko et al. (Eds.): OPTIMA 2018. Communications in Computer and Information Science. 2019. Vol. 974. P. 64–78.

[181] Vaidya P. Speeding-up linear programming using fast matrix multiplication. // 30th Annual Symposium on Foundations of Computer Science. 1989. P. 332–337.

[182] Yuyuan Ouyang, Yangyang Xu. Lower complexity bounds of firstorder methods for convex-concave bilinear saddle-point problems. // Math. Program. August 2019. URL: https://arxiv.org/pdf/1808.02901.pdf.

[183] Zhou Y., Liang Y., Shen L. A unified approach to proximal algorithms using Bregman distance. Tech. report (2016).