Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»»

На правах рукописи

Тюрин Александр Игоревич

Разработка метода решения задач структурной оптимизации

РЕЗЮМЕ
диссертации на соискание ученой степени
кандидата компьютерных наук

Москва - 2020

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики»

Научный руководитель:

Гасников Александр Владимирович, д.ф.-м.н., старший научный сотрудник Международной лаборатории стохастических алгоритмов и анализа многомерных данных, Национальный исследовательский университет «Высшая школа экономики»

1 Введение

Оптимизация играет большую роль во всех сферах человеческого общества. Сложно перечислить все направления современной деятельности, где методы и инструменты из оптимизации используется для решения различных задач. Во многих прикладных проблемах экономики, инженерии, программирования активно используется оптимизация. Появление оптимизационных методов сильно связано с зарождением вычислительной техники в 20 веке. Именно тогда и началось активное развитие современной теории оптимизации. Одним из пионеров выступал Л. Кантарович [1, 2], рассматривающий задачи линейного программирования в инженерных и экономических областях. В 50х-60х годах активно работали Г. Рубинштейн, Е. Вентцель, Н. Воробьев, Д. Юдин, Е. Гольштейн, Н. Шор, Б. Поляком. Ю. Ермольев, Л. Понтрягин и многие другие. В те года было предложено следующее: принцип максимума Понтрягина, метод проекции градиента, метод отсекающих плоскостей, штрафной функции, метод Ньютона для задачи с ограничениями, субградиентный метод, метод центров тяжести и многое другое. В 70х годах большой вклад в развитие оптимизационных методов сделали А. Немировский и Д. Юдин в работе [3]. В данной работе они использовали понятие оракул, некоторый черный ящик, который на заданную точку мог выдавать, например, значение функции и градиента. А. Немировский и Д. Юдин получили нижние оценки сходимости методов при фиксированном оракуле для некоторых достаточно общих семейств оптимизационных задач (общих выпуклых задач, липщицевых непрерывных задач, гладких строго выпуклых задач и др.). Отметим, что для некоторых из них только чуть позже были предложены оптимизационные методы, которые бы достигали данные нижние оценки. В частности, Ю. Нестеров представил алгоритм оптимизации [4], который на гладких выпуклых функциях с липщицевым градиентом имеет скорость сходимости обратно пропорциональную корню из точности решения по функции. Данная оценка является и нижней, таким образом получается, что удалось найти оптимальный метод минимизации. Аналогичный результат имеется и для сильно выпуклых гладких задач.

Получается, что для многих классов задач закрыт вопрос о построении оптимальных методов, но развитие оптимизационных методов на этом не остановилось в связи с тем, что на практике оптимизационные задачи обладают некоторой структурой, которая позволяет в теории и на практике под каждый более частный класс задач придумывать оптимизационный метод, который бы имел более быструю скорость сходимости. Отметим некоторые популярные примеры из структурной оптимизации. Композитная оптимизация решает оптимизационные задачи, в которых целевая функция представляет собой сумму гладкой и негладкой функции. Хоть и сама целевая функция негладкая, при некоторых ограниче-

ниях на негладкую часть композитной оптимизации можно предлагать методы, которые имеют скорость сходимости соответствующую гладкой оптимизации [5]. Аналогично, используя структуру задач, можно получить более оптимистичные оценки скорости сходимости для функций с гельдеровыми градиентами [6], суперпозиция функций (min-max задача) [7, 8], транспортные задачи [9, 10, 11], задача кластеризации многомерных данных в ходе избирательных компаний [12] и многие другие.

Отметим другое направление развития структурной оптимизации, связанное с различными требованиями к оракулу. В общем случае оракул — это некоторый черный ящик, который производит некоторые вычисления, сложность оптимизационных алгоритмов оценивается в количестве вызовов оракула. В классической теории оптимизации [3, 13] рассматривается оракулы, которые на запрошенную точку возвращают значение функции (оракул нулевого порядка), градиента / субградиента (оракул первого порядка), гессиана (оракул второго порядка) и так далее. В частности, для гладкой выпуклой функции f с L–липщицевым градиентом известно, что существует оракул [13] такой, что на запрошенную точку y выдает пару $(f(y), \nabla f(y))$ и выполнено неравенство

$$0 \leq f(x) - (f(y) + \langle \nabla f(y), x - y \rangle) \leq \frac{L}{2} \|x - y\|^2 \quad \forall x \in Q, \tag{1}$$

где Q — произвольное выпуклое замкнутое множество, на котором определена оптимизационная задача. Левое неравенство следует из выпуклости f, а правое неравенство следует из L–липщицевости градиента. При наличии такого оракула можно получить оптимальные оценки скорости сходимости для данного класса задача. А именно, после N вызовов оракула гарантируемая оценка скорости сходимости будет равна $\mathcal{O}\left(\frac{LR^2}{N^2}\right)$ [4, 13], где константа R есть расстояние между начальной точкой метода и ближайшей к ней оптимальной точкой. На практике и в теории у нас не всегда имеется возможность гарантировать условие (1). В случае, когда задача негладкая у нас не выполняется второе неравенство из (1). Часто на практике даже для гладких выпуклых функций с липщицевым градиентом мы не имеем возможности точно посчитать значение функции или градиента, это может возникнуть в силу неточности вычислений или в силу того, что для нахождения значения функции и градиента используется вспомогательная задача. Для таких примеров неравентсва (1) не будет выполнено, так как несложно показать, что (1) выполнено для одной пары $(f(y), \nabla f(y))$. В работе [14] представлена концепция (δ, L)–оракула, который на запрошенную точку y выдает пару $(f_\delta(y), \nabla f_\delta(y))$ и выполнено неравенство

$$0 \leq f(x) - (f_\delta(y) + \langle \nabla f_\delta(y), x - y \rangle) \leq \frac{L}{2} \|x - y\|^2 + \delta \quad \forall x \in Q. \tag{2}$$

Стоит обратить внимание, что в отличии от (1), пара $(f_\delta(y), \nabla f_\delta(y))$ не уникальна и необязательно равна $(f(y), \nabla f(y))$. Предложенный оракул

позволяет обобщать стандартные методы типа градиентного метода и быстрого градиентного на более широкий класс задач. В работе [14] в случае наличия (δ, L)–оракула для быстрого градиентного метода гарантируемая оценка скорости сходимости будет равна $\mathcal{O}\left(\frac{LR^2}{N^2} + N\delta\right)$, а для обычного градиентного метода — $\mathcal{O}\left(\frac{LR^2}{N} + \delta\right)$. Важно отметить, что в условиях получения оценок сходимости не требуется наличие гладкости функции. Наличие (δ, L)–оракула допускает достаточно широкий класс задач [14]: функции с гельдеровой гладкостью субградиента, негладкие задачи допускающие технику сглаживания [15, 16, 17], сглаживание по Моро–Иосиде [18], композитная оптимизация [5]. Отметим, что существуют и другая концепция неточного оракула [19], которая является частным случаем (δ, L)–оракула [14].

Важным свойством некоторых оптимизационных методов является их прямо–двойственность [20, 21, 22, 23], — это возможность восстанавливать достаточно эффективно решение двойственной задачи по прямой (или наоборот). Данная возможность хорошо себя зарекомендовал в транспортных задачах [9, 10, 11], задаче машинного обучения SVM и многих других [24]. Другим свойством оптимизационных методов является их возможность применения к задачам, в которых вместо настоящего градиента оракул возвращает стохастический градиент, случайный несмещенный вектор по отношению к настоящему градиенту. Методы стохастической оптимизации в последнее время являются популярными по той причине, что они позволяют уменьшать сложность подсчета градиента, что является очень важным, так как существуют примеры функций, в которых невозможно за разумное время подсчитать градиент оптимизируемой функции хотя бы в одной точке [25, 26]. Поэтому в данной работе уделяется внимание по расширению полученных результатов на прямо–двойственность и стохастическую оптимизацию.

Цели и задачи исследования: попытка унифицировать предложенные ранее методы градиентного типа в один за счет введения специальной концепции, неточной модели функции. Придумать серию методов, которые бы находили решение для обобщенных постановок оптимизационных задач и использовали структуру данных задач с помощью предложенной концепции неточной модели функции, также доказать соответствующие теоремы сходимости, по возможности, соответствующие нижним оценкам для некоторых классов задач.

Полученные результаты:

1. Предложены концепции неточной модели функции для методов градиентного типа.

2. Разработан адаптивный градиентный и быстрый градиентный метод для задач, допускающих концепцию неточной модели функции ($(\delta, L, \|\|)$–модель).

3. Разработан градиентный метод для задач с относительной гладкостью, допускающих концепцию неточной модели функции ((δ, L, V)–модель).

4. Разработан прямо-двойственный адаптивный градиентный и быстрый градиентный метод оптимизации для задач, допускающих концепцию неточной модели функции ($(\delta, L, \|\|)$–модель).

5. Разработан стохастический градиентный и быстрый градиентный метод для стохастических задач, допускающих концепцию неточной модели функции ($(\delta_1, \delta_2, L, \|\|)$–модель).

6. На основе адаптивного быстрого градиентного метода и стохастического быстрого градиентного метода предложен эвристический (без теоретических гарантий) адаптивный алгоритм для решения задач стохастической оптимизации.

Личный вклад автора включает в себя разработку оптимизационных методов для оракулов, возвращающих неточные модели функций, доказательство оценок сходимости данных методов, построение неточных моделей функции для задач из структурной оптимизации. Предложен эвристический адаптивный алгоритм оптимизации для решения задач стохастической оптимизации.

Научная новизна: предложен адаптивный градиентный и быстрый градиентный метод для оракула, возвращающего неточную модель функции, более того, предложен градиентный метод с относительной гладкостью, прямо–двойственный адаптивный градиентный и быстрый градиентный методы для неточной модели функции и стохастические неадаптивные методы. Проделана попытка по добавлению адаптивности в стохастические неадаптивные методы, в результате которой только удалось разработать эвристический адаптивный быстрый градиентный метод, показавший хорошие результаты на практике.

По теме данной диссертации было опубликовано 8 научных статей:

Публикации повышенного уровня:

1. Gasnikov A., Tyurin A. (2019) Fast gradient descent for convex minimization problems with an oracle producing a (δ, L)-model of function at the requested point. Computational Mathematics and Mathematical Physics, 59, 7, 1085–1097, Scopus Q2 (главный соавтор; автор диссертации сформулировал и доказал теорему сходимости градиентного и быстрого градиентного метода (теорема 1 и 2), представил описание примеров (раздел 4)).

2. Stonyakin F., Dvinskikh D., Dvurechensky P., Kroshnin A., Kuznetsova O., Agafonov A., Gasnikov A., Tyurin A., Uribe C., Pasechnyuk D., Artamonov S. (2019) Gradient methods for problems with inexact

model of the objective. Lecture Notes in Computer Science, 11548, 97–114, Scopus Q2 (автор диссертации подготовил текст раздела 2 и доказал теорему сходимости градиентного метода с неточной моделью с относительной гладкостью (теорема 1)).

3. Ogaltsov A., Tyurin A. (2020) A heuristic adaptive fast gradient method in stochastic optimization problems. Computational Mathematics and Mathematical Physics, 60, 7, 1108–1115, Scopus Q2 (главный соавтор; автор диссертации предложил эвристический адаптивный быстрый градиентный метод для задач стохастической оптимизации (алгоритм 2), провел его обоснование и анализ).

4. Dvurechensky P., Gasnikov A., Omelchenko A., Tyurin A. (2020) A stable alternative to Sinkhorn's algorithm for regularized optimal transport. Lecture Notes in Computer Science, 12095, 406–423, Scopus Q2 (автор диссертации помогал с разработкой алгоритма 1 и доказательством теоремы 1).

5. Dvinskikh D., Omelchenko A., Gasnikov A., Tyurin A. (2020) Accelerated gradient sliding for minimizing the sum of functions. Doklady Mathematics, 101, 3, Scopus Q2 (в печати), (автор диссертации помогал с текстом данной работы и выкладками промежуточных результатов).

Публикации стандартного уровня:

1. Tyurin A. (2020) Primal-dual fast gradient method with a model. Computer Research and Modeling, 12, 2, 263–274, Scopus Q3.

2. Dvinskikh D., Tyurin A., Gasnikov A., Omelchenko S. (2020) Accelerated and nonaccelerated stochastic gradient descent with model conception. Mathematical Notes, 108, 4, Scopus Q3 (в печати), (главный соавтор; автор диссертации разработал быстрый градиентный метод для задач стохастической оптимизации с неточной моделью функции, представил описание примеров).

3. Anikin A., Gasnikov A., Dvurechensky P., Tyurin A., Chernov A. (2017) Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints. Computational Mathematics and Mathematical Physics, 57, 8, 1262–1276, Scopus Q3. (автор диссертации помог в написании в ряде замечаний).

Доклады на конференциях и семинарах:

1. 8-я Московская международная конференция по Исследованию Операций, Россия, Москва. (17.10.2016 - 22.10.2016). Двойственный быстрый градиентный метод решения задач энтропийно–линейного программирования.

2. 59-я Всероссийская научная конференция МФТИ, Россия, Долгопрудный. (21.11.2016 - 26.11.2016). Адаптивный быстрый градиентный метод для задачи минимизации максимума выпуклых функций.

3. Воркшоп "Три оракула", Россия, Сколково. (28.12.2016). On several extensions of similar triangles method.

4. Всероссийская научная конференция с международным участием "Моделирование коэволюции природы и общества: проблемы и опыт. К 100-летию со дня рождения академика Н.Н. Моисеева", Россия, Москва. (7.11.2017 - 10.11.2017). Адаптивный метод подобных треугольников и его применение для вычисления регуляризованного оптимального транспорта

5. 60-я Всероссийская научная конференция МФТИ, Россия, Долгопрудный. (20.11.2017 - 25.11.2017). Зеркальный метод треугольника с обобщенным неточным оракулом.

6. The 23rd International Symposium on Mathematical Programming, Франция, Бордо. (1.7.2018 - 6.7.2018). Universal Nesterov's gradient method in general model conception.

7. 62-я Всероссийская научная конференция МФТИ, Россия, Долгопрудный. (18.11.2019 - 23.11.2019). Прямо-двойственный быстрый градиентный метод с моделью.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-31-90062 Аспиранты и научного проекта № 18-31-20005 мол_а_вед.

2 Постановка задачи выпуклой оптимизации

Опишем общую постановку задачи выпуклой оптимизации [13]. Пусть определена функция $f(x) : Q \longrightarrow \mathbb{R}$, множество Q является подмножеством конечномерного линейного пространства $\mathbb{R}^n$. Дана произвольная норма $\|\cdot\|$ в $\mathbb{R}^n$. Сопряженную норму определим следующим образом

$$\|\lambda\|_* = \max_{\|\nu\| \leq 1, \nu \in \mathbb{R}^n} \langle \lambda, \nu \rangle \quad \forall \lambda \in \mathbb{R}^n.$$

Введем понятие прокс–функции и дивергенции Брэгмана [27], [28] (стр. 327):

Определение 1. *$d(x) : Q \to \mathbb{R}$ называется прокс-функцией, если $d(x)$ непрерывно дифференцируемая на* int Q *и $d(x)$ является 1-сильно выпуклой относительно нормы $\|\|$ на* int Q.

Определение 2. *Функция*

$$V[y](x) = d(x) - d(y) - \langle \nabla d(y), x - y \rangle$$

называется дивергенцией Брэгмана, где $d(x)$ *— произвольная прокс–функция.*[1]

Введение дивергенции Брэгмана позволяет получать более общие результаты для оценок сходимости методов. Классическим примером дивергенции Брэгмана является функция $V[y](x) = \frac{1}{2}\|x - y\|_2^2$.

Далее будем полагать, что:

1. $Q \subseteq \mathbb{R}^n$, выпуклое, замкнутое.
2. $f(x)$ – непрерывная и выпуклая функция на Q.
3. $f(x)$ ограничена снизу на Q и достигает своего минимума в некоторой точке (необязательно единственной).

Рассматривается следующая задача оптимизации:

$$f(x) \to \min_{x \in Q}. \tag{3}$$

Решением данной выпуклой задачи будем называть такое x_*, что выполнено неравенство $f(x_*) \leq f(x)$ для любого $x \in Q$. Также, ε–решением будем называть такое x_ε, что $f(x_\varepsilon) - f(x_*) \leq \varepsilon$ для любого $x \in Q$. Основной задачей численной выпуклой оптимизации является эффективное нахождение ε–решения.

2.1 Концепция неточного решения задачи

Рассмотрим концепцию неточного решения задачи (см. [28]), которая используется в предложенных методах далее.

Определение 3. *Пусть имеется задача*

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

где $\psi(x)$ *– выпуклая, тогда* $\operatorname{Arg\,min}^{\widetilde{\delta}}_{x \in Q} \psi(x)$ *– множество таких* $\widetilde{x}$*, что*

$$\exists h \in \partial\psi(\widetilde{x}), \;\; \langle h, x - \widetilde{x} \rangle \geq -\widetilde{\delta} \;\; \forall x \in Q,$$

где $\partial\psi(\widetilde{x})$ *— субдифференциал функции* ψ *в точке* $\widetilde{x}$*. Произвольный элемент из* $\operatorname{Arg\,min}^{\widetilde{\delta}}_{x \in Q} \psi(x)$ *будем обозначать через* $\arg\min^{\widetilde{\delta}}_{x \in Q} \psi(x)$.

Данное определение является более строгим, чем определение δ–решения по функции (см. [29]). Оба определения будут эквиваленты, например, когда $\widetilde{\delta} = 0$. В некоторых более общих случаях мы можем добиться того, чтобы из δ–решения по функции получить решение в смысле определения 3 (см. [30, 29]).

[1] В работе [29] $V[y](x)$ обозначается как $V(x, y)$.

3 Содержание работы

В данном разделе рассмотрим более подробно полученные результаты и некоторые выводы.

3.1 Неточная модель функции

В работе [14] предложен (δ, L)–оракул и методы для него, которые позволяют решать большое количество оптимизационных задач. Введем понятия неточной модели ((δ, L)–модели) функции, являющиеся прямым продолжением (δ, L)–оракула.

Определение 4. *Пусть функция $\psi_\delta(x, y)$ выпуклая по x на множестве Q и выполняется условие $\psi_\delta(x, x) = 0$ для всех $x \in Q$. Будем говорить, что $\psi_\delta(x, y)$ есть $(\delta, L, \|\|)$–модель функции f в точке y относительно нормы $\|\cdot\|$ с некоторым значением $f_\delta(y)$, если для любого $x \in Q$ неравенство*

$$0 \leq f(x) - (f_\delta(y) + \psi_\delta(x, y)) \leq \frac{L}{2} \|x - y\|^2 + \delta \tag{4}$$

выполнено для некоторых $L, \delta \geq 0$.[2]

В оракульной постановке можно предполагать, что на запрошенную точку y возвращается пара $(f_\delta(y), \psi_\delta(x, y))$. Также, имеется более общее определение через относительную гладкость [32, 33, 34]:

Определение 5. *Пусть функция $\psi_\delta(x, y)$ выпуклая по x на множестве Q и выполняется условие $\psi_\delta(x, x) = 0$ для всех $x \in Q$. Будем говорить, что $\psi_\delta(x, y)$ есть (δ, L, V)–модель функции f в точке y относительно дивергенции Брэгмана V с некоторым значением $f_\delta(y)$, если для любого $x \in Q$ неравенство*

$$0 \leq f(x) - (f_\delta(y) + \psi_\delta(x, y)) \leq LV[y](x) + \delta \tag{5}$$

выполнено для некоторых $L, \delta \geq 0$.[3]

Определение 4 можно получить из определния 5, если взять дивергенцию Брэгамана $V[y](x) = \frac{1}{2} \|x - y\|^2$. Оракул, выдающий $(\delta, L, \|\|)$–модель из определения 4 является более общим, чем (δ, L)–оракул (см. (2)). В самом деле, достаточно взять $\psi_\delta(x, y) = \langle \nabla f_\delta(y), x - y \rangle$.

В работе [35] предложено обобщение определения 4 за счет введения дополнительного шума, а именно:

[2] В работах [29, 31] $(\delta, L, \|\|)$–модель обозначается как (δ, L)–модель.

[3] В работе [34] (δ, L, V)–модель обозначается как (δ, L)–модель.

Определение 6. *Пусть функция $\psi_\delta(x,y)$ выпуклая по x на множестве Q и выполняется условие $\psi_\delta(x,x) = 0$ для всех $x \in Q$. Будем говорить, что $\psi_\delta(x,y)$ есть $(\delta_1, \delta_2, L, \|\|)$–модель функции f в точке y относительно нормы $\|\cdot\|$, если для любого $x \in Q$ неравенство*

$$-\delta_1(x,y) \le f(x) - (f(y) + \psi_\delta(x,y)) \le \frac{L}{2}\|x-y\|^2 + \delta_2 \tag{6}$$

выполнено для некоторых $L \ge 0$, δ_2 и $\delta_1(x,y)$.

Можно показать, что $(\delta, L, \|\|)$–модель является $(\delta, \delta, L, \|\|)$–моделью с $\delta_1(x,y) = \delta$ и $\delta_2 = \delta$. Данная концепция полезна в задачах стохастической оптимизации (см. раздел 3.8).

3.2 Примеры неточных моделей функций

В данном разделе приведем примеры неточных моделей функции для различных оптимизационных задач.

1. *Гладкая выпуклая оптимизация с липшицевым градиентом*

   Предположим, что $f(x)$ – гладкая выпуклая функция с L-липшицевым градиентом в норме $\|\cdot\|$, тогда

   $$0 \le f(x) - f(y) - \langle \nabla f(y), x-y\rangle \le \frac{L}{2}\|x-y\|^2 \quad \forall x, y \in Q. \tag{7}$$

   Таким образом, получаем, что $\psi_{\delta_k}(x,y) = \langle \nabla f(y), x-y\rangle$ есть $(\delta, L, \|\|)$–модель функции f, $f_{\delta_k}(y) = f(y)$ и $\delta_k = 0$ для любого $k \ge 0$.

2. *Выпуклая оптимизация с гельдеровым субградиентом*

   Будем предполагать, что $f(x)$ — выпуклая функция, для которой выполняется условие Гёльдера: существует $\nu \in [0,1]$ такое, что

   $$\|\nabla f(x) - \nabla f(y)\|_* \le L_\nu \|x-y\|^\nu \quad \forall x, y \in Q.$$

   Тогда (см. [6])

   $$0 \le f(x) - f(y) - \langle \nabla f(y), x-y\rangle \le \frac{L(\delta)}{2}\|x-y\|^2 + \delta \quad \forall x, y \in Q,$$

   где

   $$L(\delta) = L_\nu \left[\frac{L_\nu}{2\delta}\frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}}$$

   и $\delta > 0$ – свободный параметр. Отсюда $\psi_{\delta_k}(x,y) = \langle \nabla f(y), x-y\rangle$ есть $(\delta, L(\delta), \|\|)$–модель функции f.

3. *Композитная оптимизация*

   Рассмотрим задачу композитной оптимизации [36]:

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

   где $g(x)$ — гладкая выпуклая функция с L-липшицевым градиентом в норме $\|\|$ и $h(x)$ — выпуклая функция (в общем случае негладкая). Для данной задачи верно следующее неравенство

$$0 \leq f(x) - f(y) - \langle \nabla g(y), x - y \rangle - h(x) + h(y) \leq \frac{L}{2} \|x - y\|^2 \qquad \forall x, y \in Q.$$

   Таким образом, мы получаем, что $\psi_{\delta_k}(x, y) = \langle \nabla g(y), x - y \rangle + h(x) - h(y)$ есть $(\delta, L(\delta), \|\|)$–модель функции f, $f_{\delta_k}(y) = f(y)$ и $\delta_k = 0$ для любого $k \geq 0$.

Отметим, что в работах [29, 34] представлено более обширное количество примеров, включающие в себя: метод условного градиента (Франк–Вульфа) [28], суперпозиция функций [7, 8], минмин задача [11], задача поиска седловой точки [11]. Для (δ, L, V)–модели в [34] имеется пример задачи о кластеризации многомерных данных, связанных с процессами, происходящими в избирательных компаниях [12].

3.3 Градиентный метод

В рамках концепции неточной модели функции в [29] были получены следующие результаты. Рассмотрим адаптивный градиентный метод для задачи (3). В алгоритме 1 будем предполагать, что дана начальная точка x_0, L_0 – константа, которая имеет смысл предположительной "локальной" константы Липшица градиента в точке x_0. Также на вход алгоритму подаются последовательности $\{\delta_k\}_{k \geq 0}$, $\{\widetilde{\delta}_k\}_{k \geq 0}$. Мы предполагаем, что на каждом шаге алгоритма k метод имеет доступ к $(\delta_k, \bar{L}_{k+1}, \|\|)$–модели. В общем случае константа $\bar{L}_{k+1}$ может меняться от итерации к итерации, мы только предполагаем, что $(\delta_k, \bar{L}_{k+1}, \|\|)$–модель существует. Мы не используем $\bar{L}_{k+1}$ константу в алгоритме 1, и, более того, нам метод адаптируется под эту константу. Последовательность $\{\widetilde{\delta}_k\}_{k \geq 0}$ – точности решения из определения 3, причем в зависимости от задачи они могут быть равными нулю, иметь постоянное значение или меняться от итерации к итерации.

Для алгоритма 1 в работе [29] получена следующая оценка сходимости.

Теорема 1 ([29])**.** *Пусть $V[x_0](x_*) \leq R^2$, где x_0 – начальная точка, а x_* – ближайшая точка минимума к точке x_0 в смысле дивергенции Брэгмана, функция f — выпуклая функция и для δ_k и точки*

Algorithm 1 Адаптивный градиентный метод с $(\delta, L, \|\|\|)$–моделью

1: **Input:** x_0 —начальная точка, $\{\delta_k\}_{k\geq 0}$, $\{\widetilde{\delta}_k\}_{k\geq 0}$ – последовательности и $L_0 > 0$.
2: $L_1 := \frac{L_0}{2}$.
3: **for** $k \geq 0$ **do**
4: Найти минимальное целое число $i_k \geq 0$ такое, что

$$f_{\delta_k}(x_{k+1}) \leq f_{\delta_k}(x_k) + \psi_{\delta_k}(x_{k+1}, x_k) + \frac{L_{k+1}}{2}\|x_k - x_{k+1}\|_2^2 + \delta_k, \quad (8)$$

где $L_{k+1} = 2^{i_k - 1}L_k$, $A_{k+1} := A_k + \frac{1}{L_{k+1}}$.

$$\phi_{k+1}(x) := \frac{1}{L_{k+1}}\psi_{\delta_k}(x, x_k) + V[x_k](x), \quad x_{k+1} := \arg\min_{x\in Q}^{\widetilde{\delta}_k}\phi_{k+1}(x).$$

5: **end for**

x_k из алгоритма 1 всегда найдется некоторая константа $\bar{L}_{k+1} > 0$ такая, что существует $(\delta_k, \bar{L}_{k+1}, \|\|\|)$–модель $\psi_{\delta_k}(\cdot, x_k)$ в точке x_k и $\bar{x}_N = \frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}x_{k+1}$. Для алгоритма 1 выполнено следующее неравенство:

$$f(\bar{x}_N) - f(x_*) \leq \frac{R^2}{A_N} + \frac{1}{A_N}\sum_{k=0}^{N-1}\widetilde{\delta}_k + \frac{2}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\delta_k.$$

Если дополнительно предположить, что на каждом шаге итерации k всегда найдется $(\delta_k, L, \|\|\|)$–модель (иначе говоря $\bar{L}_k \leq L$ для любого $k \geq 0$), то верно

$$f(\bar{x}_N) - f(x_*) \leq \frac{2LR^2}{N} + \frac{2L}{N}\sum_{k=0}^{N-1}\widetilde{\delta}_k + \frac{2}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\delta_k. \quad (9)$$

В оценке (9) в теореме 1 имеются три слагаемых соответствующие скорости сходимости, накопленным ошибкам при решении вспомогательной задачи и накопленной ошибкам от неточной модели функции. Для простоты предположим, что $\widetilde{\delta}_k = \widetilde{\delta}$ и $\delta_k = \delta$ для любого $k \geq 0$, тогда для (9) будет верно:

$$f(\bar{x}_N) - f(x_*) \leq \frac{2LR^2}{N} + 2L\widetilde{\delta} + \delta. \quad (10)$$

Из данной оценки можно сделать вывод, что данный метод имеет скорость сходимости соответствующую скорости сходимости градиентного метода [13], при этом ошибки $\widetilde{\delta}$ и δ не накапливаются по мере работы метода. В разделе 3.4 рассмотрим быстрый вариант предложенного метода, для него в оценках сходимости ошибки $\widetilde{\delta}$ и δ входить будут по-другому.

Важно отметить, что i_k в алгоритме 1 находится обычным перебором от 0 до бесконечности, но из условия о существовании $(\delta_k, L_{k+1}, \|\|)$–модели в точке x_k следует, что это этот процесс конечен, более того, несложно показать, что в среднем минимальное целое число i_k для которого выполнено (8) равно 1 (см. [6], стр. 7–8). Из этого следует, что во время работы адаптивного метода (алгоритм 1) в среднем модель $\psi_{\delta_k}(\cdot, x_k)$ запрашивается у оракула на каждой итерации 2 раза.

3.4 Быстрый градиентный метод

Рассмотрим ускоренную версию алгоритма из раздела 3.3. В работе [29] представлен алгоритм 2 и доказана соответствующая теорема 2.

Algorithm 2 Адаптивный быстрый градиентный метод с $(\delta, L, \|\|)$–моделью

1: **Input:** x_0 —начальная точка, $\{\delta_k\}_{k\geq 0}$, $\{\widetilde{\delta}_k\}_{k\geq 0}$ – последовательности и $L_0 > 0$.
2: $y_0 := x_0$, $u_0 := x_0$, $L_1 := \frac{L_0}{2}$, $\alpha_0 := 0$, $A_0 := \alpha_0$.
3: **for** $k \geq 0$ **do**
4: Найти минимальное целое число $i_k \geq 0$ такое, что

$$f_{\delta_k}(x_{k+1}) \leq f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_{k+1}, y_{k+1}) + \frac{L_{k+1}}{2}\|x_{k+1} - y_{k+1}\|^2 + \delta_k,$$

где $L_{k+1} = 2^{i_k - 1} L_k$, α_{k+1} наибольший корень $A_k + \alpha_{k+1} = L_{k+1}\alpha_{k+1}^2$, $A_{k+1} := A_k + \alpha_{k+1}$.

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

$$\phi_{k+1}(x) = V[u_k](x) + \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1}),$$
$$u_{k+1} := \arg\min_{x\in Q}^{\widetilde{\delta}_k} \phi_{k+1}(x),$$

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

5: **end for**

Теорема 2 ([29]). *Пусть $V[x_0](x_*) \leq R^2$, где x_0 – начальная точка, а x_* – ближайшая точка минимума к точке x_0 в смысле дивергенции Брэгмана, функция f — выпуклая функция и для δ_k и точки y_{k+1} из алгоритма 2 всегда найдется некоторая константа $\bar{L}_{k+1} > 0$ такая, что существует $(\delta_k, \bar{L}_{k+1}, \|\|)$–модель $\psi_{\delta_k}(\cdot, y_{k+1})$ в точке y_{k+1}. Для алго-*

ритма 2 выполнено следующее неравенство:

$$f(x_N) - f(x_*) \leq \frac{R^2}{A_N} + \frac{\sum_{k=0}^{N-1} \widetilde{\delta}_k}{A_N} + \frac{2\sum_{k=0}^{N-1} \delta_k A_{k+1}}{A_N}.$$

Если дополнительно предположить, что на каждом шаге итерации k всегда найдется $(\delta_k, L, \|\|)$–модель (иначе говоря $\bar{L}_k \leq L$ для любого $k \geq 0$), то верно

$$f(x_N) - f(x_*) \leq \frac{8LR^2}{(N+1)^2} + \frac{8L\sum_{k=0}^{N-1} \widetilde{\delta}_k}{(N+1)^2} + \frac{2\sum_{k=0}^{N-1} \delta_k A_{k+1}}{A_N}. \tag{11}$$

По аналогии с разделом 3.3 предположим, что $\widetilde{\delta}_k = \widetilde{\delta}$ и $\delta_k = \delta$ для любого $k \geq 0$, тогда для (11) будет верно:

$$f(x_N) - f(x_*) \leq \frac{8LR^2}{(N+1)^2} + \frac{8L\widetilde{\delta}}{N+1} + N\delta.$$

Сравнивая данную оценку с (10) можно сделать вывод о том, что алгоритм 2 имеет скорость сходимости быстрого градиентного метода, при этом ошибка δ по мере работы итерационного метода накапливается, а влияние ошибки $\widetilde{\delta}$ наоборот уменьшается. Если сравнивать методы относительно $\widetilde{\delta}$, то алгоритм 2 эффективнее, чем алгоритм 1. Если сравнивать методы относительно ошибки δ, то ответ не так однозначен и зависит от величины δ, более подробно данное сравнение проделано в работе [14].

По аналогии с разделом 3.3 можно сделать вывод о том, что в среднем i_k будет равно 1 [6].

3.5 Градиентный метод с относительной гладкостью

Рассмотрим упрощенный аналог метода из раздела 3.3, важной его особенностью является то, что он работает с функциями для которых имеется оракул из определения 5 с относительной гладкостью. Для многих задач (см. [33]) метод из раздела 3.3 (алгоритм 1) не может быть применим. Далее мы приводим алгоритм 3 и соответствующую теорему 3.

В данном разделе мы заменим условие на прокс–функцию $d(x)$: будем считать, что вместо 1-сильной выпуклости стоит условие только о выпуклости $d(x)$. Это позволяет применять результаты из теоремы 3 для более широкого класса задач.

Теорема 3 ([34]). *Пусть $V[x_0](x_*) \leq R^2$, где x_0 – начальная точка, а x_* – ближайшая точка минимума к точке x_0 в смысле дивергенции Брэгмана, функция f — выпуклая функция, существует (δ, L, V)–модель $\psi_\delta(\cdot, x_k)$ для f на множестве Q и $\bar{x}_N = \frac{1}{N}\sum_{k=0}^{N-1} x_{k+1}$. Для алгоритма 3 выполнено следующее неравенство:*

$$f(\bar{x}_N) - f(x_*) \leq \frac{LR^2}{N} + \widetilde{\delta} + \delta.$$

Algorithm 3 Градиентный метод с (δ, L, V)–моделью

1: **Input:** x_0 —начальная точка, $L > 0$ и $\delta, \widetilde{\delta} > 0$.
2: **for** $k \geq 0$ **do**
3:

$$\phi_{k+1}(x) := \psi_\delta(x, x_k) + LV[x_k](x), \quad x_{k+1} := \arg\min_{x \in Q}^{\widetilde{\delta}} \phi_{k+1}(x). \quad (12)$$

4: **end for**

Естественно было бы разработать ускоренный (быстрый) градиентный метод с относительной гладкостью по аналогии с разделом 3.4. Но для относительно гладких задач оценка неускоренного метода в общем случае неулучшаема с точностью до постоянной (см. [37]).

3.6 Прямо-двойственный адаптивный градиентный метод

В данном разделе рассмотрим прямо-двойственный градиентный метод, задачей которого является нахождение ε–решения не только прямой задачи (3), но и соответствующей двойственной. Введем дополнительные ограничения на множество Q, будем предполагать, что множество Q имеет следующий вид:

$$Q = \{x \mid x \in \widetilde{Q},\ f_i(x) \leq 0\ \forall i \in [1, m]\}, \quad (13)$$

где для любого i функция $f_i(x) : \widetilde{Q} \to \mathbb{R}$ выпуклая функция, и множество $\widetilde{Q}$ является выпуклым. Введем следующее обозначение:

$$F(x) = [f_1(x), \ldots, f_m(x)]^T,$$

таким образом, получаем следующую задачу оптимизации из (3):

$$f(x) \to \min_{x \in \widetilde{Q},\ F(x) \leq 0}. \quad (14)$$

Построим двойственную по Лагранжу оптимизационную задачу. Пусть

$$g(z) = \max_{x \in \widetilde{Q}} [-f(x) - \langle z, F(x) \rangle]. \quad (15)$$

Двойственной оптимизационной задачей к (14) будет задача

$$g(z) \to \min_{z \in \mathbb{R}^m_+}. \quad (16)$$

Далее будем предполагать, что выполнены условия сильной двойственности [21] (стр. 226).

Возможность восстанавливать решение двойственной задачи хорошо себя зарекомендовала, так как во многих случаях в прямой задаче можно значительно быстрее находить решение, чем в двойственной, например, это используется в транспортных задачах [9, 10, 11].

Определение 7. *Пусть x_* произвольное решение прямой задачи*

$$p(x) \to \min_{x \in \widetilde{Q},\ G(x) \leq 0}. \tag{17}$$

Точка z_* произвольное решение двойственной задачи

$$h(z) \to \min_{z \in \mathbb{R}^m_+},$$

для (17)*, где z — это двойственные переменные соответствующие ограничениям $G(x) \leq 0$. Введем оператор* argdual*, зависящий от функции $p(x)$ и $G(x)$, и возвращающий x_* и z_*:*

$$(x_*, z_*) := \operatorname*{argdual}_{x \in \widetilde{Q}}(p(x), G(x)).$$

В работе [31] предложен алгоритм 4 и получены соответствующие гарантии сходимости в теореме 4.

Algorithm 4 Прямо–двойственный адаптивный градиентный метод с $(\delta, L, \|\|)$–моделью

1: **Input:** x_0 — начальная точка, $L_0 > 0$ и $\{\delta_k\}_{k \geq 0}$.
2: $A_0 := 0$
3: **for** $k \geq 0$ **do**
4: Найти минимальное целое число $i_k \geq 0$ такое, что

$$f_{\delta_k}(x_{k+1}) \leq f_{\delta_k}(x_k) + \psi_{\delta_k}(x_{k+1}, x_k) + \frac{L_{k+1}}{2} \|x_{k+1} - x_k\|^2 + \delta_k,$$

где $L_{k+1} := 2^{i_k - 1} L_k$, $A_{k+1} := A_k + \frac{1}{L_{k+1}}$.

$$\begin{aligned} \phi_{k+1}(x) &:= \psi_{\delta_k}(x, x_k) + L_{k+1} V[x_k](x), \\ (x_{k+1}, z_{k+1}) &:= \operatorname*{argdual}_{x \in \widetilde{Q}}(\phi_{k+1}(x), F(x)). \end{aligned} \tag{18}$$

5: **end for**

Теорема 4 ([31])**.** *Пусть $\bar{x}_N = \frac{1}{A_N} \sum_{k=0}^{N-1} \frac{x_{k+1}}{L_{k+1}}$, $\bar{z}_N = \frac{1}{A_N} \sum_{k=0}^{N-1} \frac{z_{k+1}}{L_{k+1}}$, $V[x_0](x(\bar{z}_N)) \leq R^2$, где x_0 – начальная точка, а $x(\bar{z}_N)$ — точка, в которой достигается максимум в* (15) *при $z = \bar{z}_N$, функция f — выпуклая функция и для δ_k и точки x_k из алгоритма 4 всегда найдется некоторая константа $\bar{L}_{k+1} > 0$ такая, что существует $(\delta_k, \bar{L}_{k+1}, \|\|)$–модель $\psi_{\delta_k}(\cdot, x_k)$ в точке x_k. Для алгоритма 4 выполнено следующее неравенство:*

$$f(\bar{x}_N) + g(\bar{z}_N) \leq \frac{R^2}{A_N} + \frac{1}{A_N} \sum_{k=0}^{N-1} \frac{2\delta_k}{L_{k+1}}.$$

Полученные результаты полностью соответствуют результатам теоремы 1 при новых предположениях о множестве Q и условии, что $\widetilde{\delta}_k = 0$ для любого $k \geq 0$.. При этом в теореме 4 доказана оценка сходимости на зазор двойственности $f(\bar{x}_N) + g(\bar{z}_N)$.

Отметим, что существуют различные способы по восстановлению двойственного ε–решения во время нахождения ε–решения прямой задачи. В серии наших работ [20, 38] двойственные переменные по мере работы метода восстанавливаются, используя функцию Лагранжа к оптимизационной задаче (16). Такой подход приводит к тому, что возникает невязка в ограничениях. В альтернативном подходе [23] двойственные переменные восстанавливаются через вспомогательную задачу (см., например, (18)), однако, это может приводить к более плохим оценкам на зазор двойственности, так как вместо $V[x_0](x_*)$ будет стоять $V[x_0](x(\bar{z}_N))$. Методы из данного и следующего раздела базируются на подходе [23].

3.7 Прямо–двойственный адаптивный быстрый градиентный метод

В данном разделе рассматривается быстрый вариант метода из раздела 3.6. Будем налагать те же ограничения на множество Q, что и в разделе 3.6. В работе [31] предложен алгоритм 5 и получены соответствующие гарантии сходимости в теореме 5.

Теорема 5 ([31]). *Пусть $\bar{z}_N = \frac{1}{A_N}\sum_{k=0}^{N-1} \alpha_{k+1} z_{k+1}$, $V[x_0](x(\bar{z}_N)) \leq R^2$, где x_0 – начальная точка, а $x(\bar{z}_N)$ — точка, в которой достигается максимум в (15) при $z = \bar{z}_N$, функция f — выпуклая функция и для δ_k и точки y_{k+1} из алгоритма 5 всегда найдется некоторая константа $\bar{L}_{k+1} > 0$ такая, что существует $(\delta_k, \bar{L}_{k+1}, \|\|)$–модель $\psi_{\delta_k}(\cdot, y_{k+1})$ в точке y_{k+1}. Для алгоритма 5 выполнено следующее неравенство:*

$$f(x_N) + g(\bar{z}_N) \leq \frac{R^2}{A_N} + \frac{2}{A_N}\sum_{k=0}^{N-1} A_{k+1}\delta_k.$$

С учетом ограничения (13) на множество Q полученная оценка скорости сходимомти соответствует оценки скорости сходимости из теоремы 2 при условии, что $\widetilde{\delta}_k = 0$ для любого $k \geq 0$.

3.8 Стохастический быстрый градиентный метод

В данном разделе рассматривается $(\delta_1, \delta_2, L, \|\|)$–модель из определения 6, которая является обобщением $(\delta, L, \|\|)$–модели функции. По аналогии с разделом 3.3 и 3.4 в работе [35] представлены оценки сходимости для методов, использующих $(\delta_1, \delta_2, L, \|\|)$–модель. Одним из самых важных следствий является то, что данная модель очень удачно подходит для

Algorithm 5 Прямо–двойственный адаптивный быстрый градиентный метод с $(\delta, L, \|\|\|)$–моделью

Input: x_0 —начальная точка, $\{\delta_k\}_{k\geq 0}$ и $L_0 > 0$.

$y_0 := x_0$, $u_0 := x_0$, $L_1 := \frac{L_0}{2}$, $\alpha_0 := 0$, $A_0 := \alpha_0$.

for $k \geq 0$ **do**

Найти минимальное целое число $i_k \geq 0$ такое, что

$$f_{\delta_k}(x_{k+1}) \leq f_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_{k+1}, y_{k+1}) + \frac{L_{k+1}}{2}\|x_{k+1} - y_{k+1}\|^2 + \delta_k,$$

где $L_{k+1} = 2^{i_k - 1}L_k$, α_{k+1} наибольший корень $A_k + \alpha_{k+1} = L_{k+1}\alpha_{k+1}^2$, $A_{k+1} := A_k + \alpha_{k+1}$.

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

$$\phi_{k+1}(x) := \psi_{\delta_k}(x, y_{k+1}) + L_{k+1}V[u_k](x),$$
$$(x_{k+1}, z_{k+1}) := \operatorname*{argdual}_{x \in \widetilde{Q}}(\phi_{k+1}(x), F(x)).$$

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

end for

задач стохастической оптимизации [39, 40]. В алгоритме 6 представлен быстрый градиентный метод для $(\delta_1, \delta_2, L, \|\|_2)$–модели.

Algorithm 6 Быстрый градиентный метод с $(\delta_1, \delta_2, L, \|\|_2)$–моделью

1: **Input:** x_0 —начальная точка и $L > 0$.
2: $y_0 := x_0$, $u_0 := x_0$, $\alpha_0 := 0$, $A_0 := \alpha_0$.
3: **for** $k \geq 0$ **do**
4: Константа α_{k+1} — это наибольший корень $A_k + \alpha_{k+1} = L\alpha_{k+1}^2$, $A_{k+1} := A_k + \alpha_{k+1}$.

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

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

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

5: **end for**

Далее для $(\delta_1, \delta_2, L, \|\|)$–модели сформулируем теорему сходимости.

Теорема 6 ([35]). *Пусть $\frac{1}{2}\|x_* - x_0\|_2^2 \leq R^2$, где x_0 — начальная точка, а x_* — ближайшая точка минимума к точке x_0 в смысле евклидова расстояния, функция f — выпуклая функция, существует $(\delta_1^k, \delta_2^k, L, \|\|_2)$–модель $\psi_{\delta_k}(\cdot, y_{k+1})$ в точке y_{k+1} для функции f, где y_{k+1} — точка из алгоритма 6. Для алгоритма 6 выполнено следующее неравенство:*

$$\begin{aligned} f(x_N) - f(x_*) \leq \frac{4LR^2}{N^2} + \frac{1}{A_N}\sum_{k=0}^{N-1} A_k \delta_1^k(x_k, y_{k+1}) \\ + \frac{1}{A_N}\sum_{k=0}^{N-1} \alpha_{k+1}\delta_1^k(x_*, y_{k+1}) + \frac{1}{A_N}\sum_{k=0}^{N-1} A_{k+1}\delta_2^k. \end{aligned} \tag{19}$$

Если дополнительно предполагать, что: $\{\delta_1^k\}_{k=0}^{N-1}$ и $\{\delta_2^k\}_{k=0}^{N-1}$ случайные последовательности для которых выполнено:

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

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

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

Предположение 2. *Пусть даны две последовательности* $\delta_1^k(x,y)$ *и* δ_2^k *(*$k \geq 0$*). Случайная величина* $\delta_1^k(x,y)$ *имеет* $\left(\hat{\delta}_1^k(x-y)\right)^2$*-субгауссовский условный момент (*$\hat{\delta}_1^k(\cdot)$ *есть неслучайная функция от одного аргумента) такой, что*

1. $\hat{\delta}_1^k(\alpha z) \leq \alpha \hat{\delta}_1^k(z)$ *для всех* $\alpha \geq 0$ *и* $z \in B(0,R)$.
2. $\hat{\delta}_1 < +\infty$*, где* $\hat{\delta}_1 \geq \sup_{z \in B(0,R)} \hat{\delta}_1^k(z)$.

Тогда с большой вероятностью [4]

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

причем

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

Важным следствием является стохастическая оптимизация. Пусть дана следующая задача оптимизации:

$$f(x) = \mathbb{E}[f(x,\xi)] \to \min_{x \in Q}, \tag{20}$$

где множество Q предполагается выпуклым и замкнутым, ξ — случайная величина, математическое ожидание $\mathbb{E}[f(x,\xi)]$ определено и конечно для любого $x \in Q$, функция f — выпуклая и имеет L–Липшицев градиент, $\nabla f(y,\xi)$ имеет субгауссовое распределение с субгауссовской дисперсией σ^2. Для задачи (20) можно взять модель $\psi_\delta(x,y) = \langle \nabla f(y,\xi), x-y \rangle$, при этом несложно показать (см. [35]), что для $\{\delta_k^1\}_{k=0}^{N-1}$ и $\{\delta_k^2\}_{k=0}^{N-1}$ верно: $\hat{\delta}_1 = O(\sigma R)$ и $\hat{\delta}_2 = O(\sigma^2/L)$. Далее, используя технику mini–batch (см. [35]), можно получить оптимальные оценки сходимости для задачи (20).

Важно отметить, что подобные рассуждения можно проводить и для $(\delta_1, \delta_2, L, \|\|)$–моделей, предусмотренных, например, для композитной оптимизации или min–max задачи.

3.9 Эвристический адаптивный стохастический быстрый градиентный метод

На основе адаптивного быстрого градиентного метода (алгоритм 2) и неадаптивного стохастического быстрого градиентного метода [35] (алгоритм 6) был предложен эвристический адаптивный стохастический быстрый градиентный метод в работе [41]. На данный момент нам неизвестно, можно ли добавить адаптивность к быстрому стохастическому

[4] означает с вероятностью $\geq 1-\gamma$, а $\tilde{O}(\cdot)$ означает то же самое, что $O(\cdot)$, только числовой множитель зависит от $\ln(1/\gamma)$.

градиенту таким образом, чтобы оценки скорости сходимости сохранились, различные попытки проделаны в работах [42, 43, 44, 45, 46, 47], более подробный анализ можно найти в [41]. Обозначим mini–batch для градиентов, как

$$\widetilde{\nabla}^{m_{k+1}} f(y) = \frac{1}{m_{k+1}} \sum_{j=1}^{m_{k+1}} \nabla f(y; \xi_j),$$

а mini–batch для значений функций, как

$$f^{m_{k+1}}(y) = \frac{1}{m_{k+1}} \sum_{j=1}^{m_{k+1}} f(y; \xi_j),$$

где ξ_j —случайная величина ($j = 1, ..., m_{k+1}$), $\nabla f(y; \xi_j)$ и $f(y; \xi_j)$ — несмещенные оценки $\nabla f(y)$ и $f(y)$, m_{k+1} — количество элементов в mini–batch. В алгоритме 7 представлен эвристический метод из [41], который вместо значений функции и градиентов использует их стохастические аппроксимации. В работе [35] на основе модели из определения 6 представлено доказательство сходимости для неадаптивного варианта алгоритма 7.

Отметим, что в работе [41] была проделана серия экспериментов на практических задачах машинного обучения MNIST [48] и CIFAR [49] и показано, что предложенный подход на логистической регрессии с линейной, нейросетевой, сверточно–нейросетевой моделями показывал лучше результаты, чем популярные оптимизационные методы Adam [50] и AdaGrad [51].

4 Заключение

В результате подготовки данной диссертации были опубликованы статьи [29, 34, 31, 35, 41, 38, 20, 52].

В статьях [29, 34, 31, 35] разработаны методы оптимизации вокруг концепции неточной модели функции и доказаны оценки сходимости этих методов. В данных работах приведено большое количество постановок задач, которые вписываются в концепцию неточной модели функции. Кроме стандартных задач структурной оптимизации, таких как гладкая оптимизация, композитная оптимизация, оптимизация функция с гельдеровыми градиентами, предложенная концепция описывает многие задачи: транспортные задачи [29, 11], задачи о кластеризации многомерных данных, связанных с процессами, происходящими в избирательных компаниях [34, 12] и многие другие.

Работы [41, 38, 20, 52] являются отправными точками для развития концепции неточной модели функции и мотивирующими для дальнейших исследований.

Algorithm 7 Эвристический адаптивный быстрый стохастический градиентный метод

Input: x_0 —начальная точка, константы $\epsilon > 0$, $L_0 > 0$ и $\sigma_0^2 > 0$.

$y_0 := x_0$, $u_0 := x_0$, $L_1 := \frac{L_0}{2}$, $\alpha_0 := 0$, $A_0 := \alpha_0$.

for $k \geq 0$ **do**

Найти минимальное целое число $i_k \geq 0$ такое, что

$$f^{m_{k+1}}(x_{k+1}) \leq f^{m_{k+1}}(y_{k+1}) + \langle \widetilde{\nabla}^{m_{k+1}} f(y_{k+1}), x_{k+1} - y_{k+1} \rangle + \\ + \frac{L_{k+1}}{2} \|x_{k+1} - y_{k+1}\|^2 + \frac{\epsilon}{L_{k+1}\alpha_{k+1}},$$

где $L_{k+1} = 2^{i_k - 1} L_k$, $\widetilde{\alpha}_{k+1}$ наибольший корень $A_k + \alpha_{k+1} = L_k \alpha_{k+1}^2$, α_{k+1} наибольший корень $A_k + \alpha_{k+1} = L_{k+1}\alpha_{k+1}^2$, $A_{k+1} := A_k + \alpha_{k+1}$, $m_{k+1} := \left\lceil \frac{3\sigma_0^2 \tilde{\alpha}_{k+1}}{\epsilon} \right\rceil$. Если $i_k = 0$, то сгенерировать i.i.d. ξ_j ($j = 1, ..., m_{k+1}$).

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

$$\phi_{k+1}(x) := \frac{1}{2}\|x - u_k\|_2^2 + \alpha_{k+1}\left(f^{m_{k+1}}(y_{k+1}) + \langle \widetilde{\nabla}^{m_{k+1}} f(y_{k+1}), x - y_{k+1} \rangle\right),$$

$$u_{k+1} := \arg\min_{x \in Q} \phi_{k+1}(x),$$

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

end for

Перечислим основные полученные результаты, которые достигнуты в данной диссертации и выносятся на защиту:

1. Разработаны различные концепции неточной модели функции для методов градиентного типа. Показано, что данные концепции описывают большое количество современных оптимизационных задач.

2. Предложены различные оптимизационные методы, допускающие концепции неточной модели функции. Для методов из разделов 3.3–3.8 доказаны теоремы сходимости и проведен анализ.

3. Разработан эвристический адаптивный стохастический быстрый градиентный метод оптимизации и проделано его обоснование.

4. Проделан теоретический анализ вокруг прямо–двойственных методов для задач с сильно выпуклыми функционалами простой структуры при аффинных ограничениях, задачи по вычислению регуляризованного оптимального транспорта и задач с неточной моделью функции.

Стоит отметить, что в данную диссертацию не вошли некоторые исследования и примеры постановок задач, для которых полезна концепция неточной модели функции, которые планируется добавить в наших следующих работах:

1. В следующих статьях планируется все результаты перенести на задачи с сильно выпуклыми функционалами [30].

2. Нами в работе [52] рассматривается практически значимый вид задачи, имеющий вид суммы гладких сильно выпуклых функций с гладким регуляризатором. В данной работе предлагается подход, позволяющий получать оптимальные оценки для случая, когда композитный член не является проксимально дружественным[5]. В будущих исследованиях планируется перенести данный результат на модельную общность.

3. Планируется распространить концепцию неточной модели функции на блочно–компонентную оптимизацию [53, 54]. Как и в общей стохастической оптимизации, блочно–компонентные методы позволяют эффективно делать шаг метода за счет того, что направление спуска оценивается не по всем координатам оптимизируемой переменной, а только по некоторому подмножеству (блоку) координат.

[5]проксимальная дружественность функции — это возможность находить минимум данной функции с квадратичной добавкой пренебрежительно быстро.

4. В конце раздела 3.6 упоминается подход по восстановлению двойственных переменных, описанный в работах [20, 38]. Было бы полезно рассмотреть возможность по обобщению методов из [20, 38] с помощью концепции неточной модели функции.

5 Список литературы

[1] *Kantorovich L.* Mathematical methods of organizing and planning production // Management Science. 1960. V. 6, №. 4. P. 366–422.

[2] *Polyak B.* History of mathematical programming in the USSR: analyzing the phenomenon // Math. Program. 2002. V. 91. № 3. P. 401–416.

[3] *Nemirovski A., Yudin D.* Problem complexity and method efficiency in optimization. Wiley–Interscience. 1983.

[4] *Nesterov Yu.* A method of solving a convex programming problem with convergence rate $O(1/k^2)$ // Dokl. Akad. Nauk SSSR. 1983. V. 269. № 3. P. 543–547.

[5] *Nesterov Yu.* Gradient methods for minimizing composite functions // Math. Program. 2013. V. 140, № 1. P. 125–161.

[6] *Nesterov Yu.* Universal gradient methods for convex optimization problems // Math. Program. 2015. V. 152. № 1–2. P. 381–404.

[7] *Nemirovski A.* Information-based complexity of convex programming. Technion. 1995.

[8] *Lan G.* Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization // Math. Program. 2015. V. 149. № 1–2. P. 1–45.

[9] *Baimurzina D., Gasnikov A., Gasnikova E., Kubentaeva M., Lagunovskaya A., Dvurechensky P., Ershov E.* Universal Method of Searching for Equilibria and Stochastic Equilibria in Transportation Networks // Computational Mathematics and Mathematical Physics. 2019. V. 59. № 1. P. 19–33.

[10] *Gasnikov A., Gasnikova E., Nesterov Yu.* Dual methods for finding equilibriums in mixed models of flow distribution in large transportation networks // Computational Mathematics and Mathematical Physics. 2017. V. 58. № 9. P. 1395–1403.

[11] *Gasnikov A.* Effective numerical methods for finding equilibrium in large transport networks. PhD thesis. MFTI, 2016.

[12] *Nesterov Yu.* Soft clustering by convex electoral model // CORE Discussion paper. 2018/01. 20p. URL: https://alfresco.uclouvain.be/alfresco/service/guest/streamDownload/workspace/SpacesStore/ff42ec88-4339-4223-b05d-b768c71ef4e6/coredp2018_01web.pdf?guest=true.

[13] *Nesterov Yu.* Lectures on convex optimization. Springer. 2018.

[14] *Devolder O., Glineur F., Nesterov Yu.* First-order methods of smooth convex optimization with inexact oracle // Math. Program. 2014. V. 146. № 1–2. P. 37–75.

[15] *Nesterov Yu.* Smooth minimization of non-smooth functions // Math. Program. 2005. V. 103. № 1. P. 127–152.

[16] *Nesterov Yu.* Excessive gap technique in nonsmooth convex minimization // SIAM J. Optimizat. 2005. V. 16, № 1. P. 235–249.

[17] *Nesterov Yu.* Smoothing technique and its applications in semidefinite optimization // Math. Program. 2007. V. 110. № 2. P. 245–259.

[18] *Lemarechal C., Sagaztizabal C.* Practice aspects of Moreau–Yosida regularization: Theoretical preliminaries // SIAM J. Optimizat. 1997. V. 7, № 2. P. 367–385.

[19] *D'Aspremont A,* Smooth optimization with approximate gradient // SIAM J. Optimizat. 2019. V. 19, № 3. P. 1171–1183.

[20] *Anikin A., Gasnikov A., Dvurechensky P., Tyurin A., Chernov A.* Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints // Computational Mathematics and Mathematical Physics. 2017. V. 57. № 8. P. 1262–1276.

[21] *Boyd S., Vandenberghe L.* Convex optimization. Cambridge University Press. 2004.

[22] *Nesterov Yu.* Complexity bounds for primal-dual methods minimizing the model of objective function // Mathematical Programming. 2018. V. 171, №. 1–2. P. 311–330.

[23] *Nesterov Yu.* Primal-dual subgradient methods for convex problems. // Mathematical Programming. 2009. V. 120, № 1. P. 221–259.

[24] *Gasnikov A.* Universal gradient descent // e-print. arXiv:1711.00394. 2020.

[25] *Goodfellow I., Bengio Y., Courville A.* Deep learning. MIT press. 2016.

[26] *Krizhevsky A., Sutskever I., Hinton G.* Imagenet classification with deep convolutional neural networks // In Advances in neural information processing systems. 2012. P. 1097–1105.

[27] *Bregman L.* The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming // USSR Computational Mathematics and Mathematical Physics. 1967. V. 7. № 3. P. 200–217.

[28] *Ben-Tal A., Nemirovski A.* Lectures on Modern Convex Optimization. Philadelphia: SIAM, 2015. URL: `http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf`.

[29] *Gasnikov A., Tyurin A.* Fast gradient descent for convex minimization problems with an oracle producing a (δ, L)-model of function at the requested point // Computational Mathematics and Mathematical Physics. 2019. V. 59. № 7. P. 1085–1097.

[30] *Stonyakin F., Tyurin A., Gasnikov A., Dvurechensky P., Agafonov A., Dvinskikh D., Pasechnyuk D., Artamonov S., Piskunova V.* Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model // e-print. arXiv:2001.09013. 2020.

[31] *Tyurin A.* Primal–dual fast gradient method with a model // Computer Research and Modeling. 2020. V. 12, № 2. P. 263–274.

[32] *Bauschke H., Bolte J., Teboulle M.* A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications // Mathematics of Operations Research. 2016. V. 42. № 2. P. 330–348.

[33] *Lu H., Freund R., Nesterov Yu.* Relatively smooth convex optimization by first-order methods, and applications // SIAM J. Optimizat. 2018. V. 28, № 1. P. 333–354.

[34] *Stonyakin F., Dvinskikh D., Dvurechensky P., Kroshnin A., Kuznetsova O., Agafonov A., Gasnikov A., Tyurin A., Uribe C., Pasechnyuk D., Artamonov S.* Gradient methods for problems with inexact model of the objective // Lecture Notes in Computer Science. 2019. V. 11548. P. 97–114.

[35] *Dvinskikh D., Tyurin A., Gasnikov A., Omelchenko S.* Accelerated and nonaccelerated stochastic gradient descent with model conception // Mathematical Notes. 2020. V. 108. № 4. In press.

[36] *Nesterov Yu.* Gradient methods for minimizing composite functions // Math. Program. 2013. V. 140, № 1. P. 125–161.

[37] *Dragomir R., Taylor A., D'Aspremont A., Bolte J.* Optimal complexity and certification of Bregman first-order methods // e-print. arXiv: 1911.08510. 2019.

[38] *Dvurechensky P., Gasnikov A., Omelchenko A., Tyurin A.* A stable alternative to Sinkhorn's algorithm for regularized optimal transport // Lecture Notes in Computer Science. 2020. V. 12095. P. 406–423.

[39] *Lan G.* Lectures on optimization. Methods for Machine Learning // e-print. 2019. URL: http://pwp.gatech.edu/guanghui-lan/wp-content/uploads/sites/330/2019/08/LectureOPTML.pdf.

[40] *Devolder O.* Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. PhD thesis. CORE UCL, 2013.

[41] *Ogaltsov A., Tyurin A.* A heuristic adaptive fast gradient method in stochastic optimization problems // Computational Mathematics and Mathematical Physics. 2020. V. 60. № 7. P. 1108–1115.

[42] *Bach F., Levy K.Y.* A universal algorithm for variational inequalities adaptive to smoothness and noise // COLT, 2019.

[43] *Vaswani S., Mishkin A., Laradji I., Schmidt M., Gidel G., Lacoste-Julien S.* Painless Stochastic Gradient: interpolation, line-search, and convergence rates // NIPS, 2019.

[44] *Ward R., Wu X., Bottou L.* AdaGrad stepsizes: sharp convergence over nonconvex landscapes, from any initialization // ICML, 2019.

[45] *Deng Q., Cheng Y., Lan G.* Optimal adaptive and accelerated stochastic gradient descent // e-print. arXiv:1810.00553. 2018.

[46] *Levy K.Y., Yurtsever A., Cevher V.* Online adaptive methods, universality and acceleration // NIPS, 2018.

[47] *Iusem A.N., Jofre A., Oliveira R.I., Thompson P.* Variance-based extragradient methods with line search for stochastic variational inequalities // SIAM J. Optimizat. 2019. V. 29, № 1. P. 175–206.

[48] *LeCun Y., Bottou L., Bengio Y., Haffner P.* Gradient-based learning applied to document recognition // Proceedings of the IEEE. 1998. V. 86. № 11. P. 2278–2324.

[49] *Krizhevsky A.* Learning Multiple Layers of Features from Tiny Images. PhD thesis. University of Toronto, 2009.

[50] *Kingma D.P., Ba J.* Adam: a method for stochastic optimization // ICLR, 2015.

[51] *Duchi J., Hazan E., Singer Y.* Adaptive subgradient methods for online learning and stochastic optimization // Journal of Machine Learning Research. 2011. V. 12. № Jul. P. 2121–2159.

[52] *Dvinskikh D., Omelchenko A., Gasnikov A., Tyurin A.* Accelerated gradient sliding for minimizing the sum of functions // Doklady Mathematics. 2020. V. 101. № 3. In press.

[53] *Dvurechensky P., Gasnikov A., Tyurin A.* Randomized similar triangles method: a unifying framework for accelerated randomized optimization methods (coordinate descent, directional search, derivative–free method) // e-print. arXiv: 1707.08486. 2017.

[54] *Nesterov Yu.* Efficiency of coordinate descent methods on huge–scale optimization problems // SIAM J. Optimizat. 2012. V. 22, № 2. P. 341–362.