Федеральное государственное автономное образовательное учреждение
высшего образования «Московский физико-технический институт
(национальный исследовательский университет)»

*На правах рукописи*

Двуреченский Павел Евгеньевич

# Численные методы оптимизации для задач большой размерности: неточный оракул и прямо-двойственный анализ

РЕЗЮМЕ
диссертации на соискание ученой степени
доктора компьютерных наук

Москва - 2020

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Московский физико-технический институт (национальный исследовательский университет)».

**Научный консультант:**

Гасников Александр Владимирович, д.ф.-м.н., доцент кафедры математических основ управления, Федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)».

# 1 Введение

Численные методы оптимизации остаются активной областью исследований с 1980-х годов, мотивированной широким спектром приложений, например в исследовании операций, оптимальном управлении. Начиная с работ [1, 2], одним из основных направлений исследований стали методы внутренней точки. Эти методы сочетают шаги метода Ньютона с идее штрафных функций и позволяют решать очень общий класс выпуклых задач за полиномиальное время, что подтверждено как теоретическими результатами, так и практической эффективностью. Новый век поставил перед численными методами оптимизации новые задачи. Благодаря увеличению объема доступных данных и более мощным вычислительным ресурсам, машинное обучение стало областью интенсивных исследований. Типичной задачей оптимизации в машинном обучении является минимизация эмпирического риска, ключевой особенностью которой является большое число переменных и большое количество компонент в целевой функции, которая является суммой индивидуальных функций потерь. В этом случае итерация метода Ньютона становится дорогой, поскольку требует обращения матрицы. Это мотивировало пожертвовать логарифмической зависимостью от точности в пользу дешевой итерации и использования методов первого порядка для решения таких задач. Другая причина заключается в том, что данные обычно зашумлены, и нет необходимости решать задачу с высокой точностью в приложении к машинному обучению. Другими важными приложениями методов первого порядка является обработка сигналов и анализ изображений, где цель состоит в том, чтобы восстановить многомерный сигнал по данных высокой размерности, например восстановить зашумленное изображение.

Перечисленные приложения подтолкнули уже давно известные [3, 4, 5] методы первого порядка к новой фазе своего развития в 2000-х годах. Некоторые важные факты об этих методах были известны уже 15 лет. В частности, концепция оракула типа черный ящик [6] позволила получить нижние оценки сложности по наихудшему случаю для различных классов задач и методов. В частности, был обнаружен зазор между нижней границей $O(1/k^2)$ и верхней границей $O(1/k)$ скорости сходимости градиентного метода для минимизации выпуклых гладких функций. Здесь $k$ - число итераций. Этот зазор привел к открытию важноого свойства ускорения для методов первого порядка и к созданию ускоренного градиентного метода [7]. В новом столетии были предложены многие обобщения этого алгоритма, мотивированные задачами обработки изображений и машинного обучения, включая композитные версии [8, 9], ускоренный метод стохастического градиента [10], ускоренные методы с редукцией дисперсии [11, 12, 13, 14, 15]. В дополнение к ускоренным методам стохастического градиента для задач минимизации конечной сум-

мы, которые используют случайный выбор градиента компоненты, ускорение было получено для других рандомизированных методов, таких как случайный покоординатный спуск [16] и рандомизированные градиентные методы [17]. Последнее мотивировано задачами, в которых доступен только оракул нулевого порядка, например когда целевая функция сама по себе является решением некоторой вспомогательной задачи. Для этой постановки важно проанализировать методы нулевого порядка с неточными значениями функции, поскольку эту вспомогательную задачу можно решить только неточно. При использовании методов первого порядка на практике также можно встретить неточность в градиенте. Ускоренный градиентный метод был проанализирован в [18], а важная конструкция неточного оракула первого порядка была представлена в [19]. Еще одно важное обобщение ускоренных методов первого порядка - это ускоренные методы для задач с линейными ограничениями, которые были предложены в [20], но с неоптимальной скоростью $O(1/k)$ для невязки по ограничениям.

**Цели и задачи исследования.** Цель диссертации состоит из двух частей. Первая цель - дальнейшее обобщение существующих методов первого и нулевого порядка для задач с неточностями в функциях и значениях градиента, причем неточности являются детерминированными или стохастическими. Вторая цель - предложить новые прямодвойственные методы первого порядка, которые позволяют одновременно решать прямую и двойственную задачи с оптимальной скоростью сходимости. Особое внимание уделяется задачам с линейными ограничениями и применению предложенных методов к задачам вычисления оптимального транспортного расстояния и барицентра.

**Полученные результаты:**

1. Предложен стохастический промежуточный градиентный метод для выпуклых задач со стохастическим неточным оракулом.

2. Предложен градиентный метод с неточным оракулом для детерминированной невыпуклой оптимизации.

3. Предложен безградиентный метод с неточным оракулом для детерминированной выпуклой оптимизации.

4. Предложен метод для вычисления производной вектора ранжирования веб-страниц и в сочетании с двумя вышеупомянутыми методами предложены методы оптимизации нулевого и первого порядка для обучения модели ранжирования веб-страниц.

5. Предложена концепция неточного оракула для методов, использующих производные по направлениям, и предложен ускоренный метод, использующий производные по направлению, для гладкой

стохастической выпуклой оптимизации. Также предложен ускоренный и неускоренный метод, использующий производные по направлению, для сильно выпуклой гладкой стохастической оптимизации.

6. Предложены прямодвойственные методы поиска седловых точек в бесконечномерных играх (дифференциальных играх) в выпукло-вогнутой и сильно выпукло-вогнутой постановках.

7. Предложены неадаптивный и адаптивный ускоренный прямодвойственный метод для задач сильно выпуклой минимизации с линейными ограничениями типа равенства и неравенства.

8. Этот алгоритм применен к задаче оптимального транспорта и получены новые оценки сложности для этой задачи, которые в некоторых режимах лучше, чем оценки для алгоритма Синхорна.

9. Предложен стохастический ускоренный прямодвойственный метод для задач с линейными ограничениями, который применен к задаче аппроксимации барицентра Васерштейна.

10. Предложено прямодвойственное обобщение ускоренных методов, которые используют линейный поиск для определения размера шага и адаптации к константе Липшица градиента.

**Личный вклад автора** включает в себя разработку перечисленных выше методов оптимизации, доказательство теорем о скорости сходимости и сложности для этих методов, а также приложения этих методов к задачам оптимального транспорта и задаче обучения для модели ранжирования веб-страниц.

**Научная новизна:** Предлагаемые варианты ускоренных методов первого и нулевого порядка для выпуклой оптимизации при различных типах неточностей являются новыми. Предлагаемые прямодвойственные методы для перечисленных постановок задач также являются новыми и позволяют получить новые методы решения оптимальных транспортных задач. В частности, получены новые результаты о сложности для нерегуляризованной оптимальной транспортной задачи и новый распределенный алгоритм для аппроксимации барицентра Вассерштейна набора мер с использованием выборок из этих мер.

По теме данной диссертации было опубликовано 10 научных статей.

**Публикации повышенного уровня:**

1. Dvurechensky, P., and Gasnikov, A. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. Journal of Optimization Theory and Applications 171, 1 (2016), 121–145, Scopus Q1 (главный соавтор; автор диссертации разработал основные алгоритмы, сформулировал и доказал теоремы о скорости сходимости предложенных методов).

2. Gasnikov, A. V., and Dvurechensky, P. E. Stochastic intermediate gradient method for convex optimization problems. Doklady Mathematics 93, 2 (2016), 148–151, Scopus Q2 (главный соавтор; автор диссертации разработал основные алгоритмы, сформулировал и доказал теоремы о скорости сходимости предложенных методов).

3. Bogolubsky, L., Dvurechensky, P., Gasnikov, A., Gusev, G., Nesterov, Y., Raigorodskii, A. M., Tikhonov, A., and Zhukovskii, M. Learning supervised pagerank with gradient-based and gradient-free optimization methods. In Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 4914–4922, CORE A* (автор диссертации разработал общий безградиентный (Algorithm 1,2) и градиентный (Algorithm 3,4) методы с неточным оракулом, предложил алгоритм для приближенного вычисления производной вектора ранжирования веб-страниц, сформулировал и доказал теоремы о скорости сходимости для предложенных методов: Lemma 1,2, Theorem 1-4).

4. Dvurechensky, P., Gorbunov, E., and Gasnikov, A. An accelerated directional derivative method for smooth stochastic convex optimization. European Journal of Operational Research (2020), `https://doi.org/10.1016/j.ejor.2020.08.027`, Scopus Q1 (главный соавтор; автор диссертации предложил концепцию неточного оракула, использующего производные по направлению, в стохастической выпуклой оптимизации, доказал (в неразрывном сотрудничестве с Э.Горбуновым) теорему о скорости сходимости (Theorem 1) для ускоренного спуска по направлению, доказал теоремы сходимости (Theorems 3,4) для сильно выпуклых задач).

5. Dvurechensky, P., Nesterov, Y., and Spokoiny, V. Primal-dual methods for solving in infinite-dimensional games. Journal of Optimization Theory and Applications 166, 1 (2015), 23–51, Scopus Q1 (главный соавтор; автор диссертации предложил основные алгоритмы и доказал теоремы об их скорости сходимости).

6. Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In Proceedings of the 5th International Conference on Machine Learning (2018), J. Dy and A. Krause, Eds., vol. 80 of Proceedings of Machine Learning Research, pp. 1367–1376, CORE A* (главный соавтор; автор диссертации предложил общий прямодвойственный адаптивный ускоренный градиентный метод (Algorithm 3) для задач с линейными ограничениями, доказал теорему о скорости сходимости (Theorem 3), предложил алгоритм для аппрокси-

мации оптимального транспортного расстояния (Algorithm 4), получил оценку сложности для этого алгоритма (Theorem 4), выполнил численные эксперименты для сравнения этого метода и метода Синхорна).

7. Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C. A., and Nedić, A. Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In Advances in Neural Information Processing Systems 31 (2018), S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., NeurIPS 2018, Curran Associates, Inc., pp. 10783–10793, CORE A* (главный соавтор; автор диссертации предложил общую идею статьи, общий прямодвойственный ускоренный стохастический градиентный метод (Algorithm 2) для задач с линейными ограничениями, доказал Theorem 2 о его скорости сходимости, предложил алгоритм для вычисления барицентра Васерштейна (Algorithm 4), доказал (в неразрывном сотрудничестве с Д. Двинских) теорему о его сложности).

8. Guminov, S. V., Nesterov, Y. E., Dvurechensky, P. E., and Gasnikov, A. V. Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems. Doklady Mathematics 99, 2 (2019), 125-128, Scopus Q2 (автор диссертации предложил прямо-двойственный вариант ускоренного градиентного метода с линейным поиском для задач с линейными ограничениями, доказал теорему о его сходимости (Theorem 3)).

9. Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. Optimization Methods and Software (2020), `https://doi.org/10.1080/10556788.2020.1731747`, Scopus Q1 (автор диссертации предложил прямо-двойственный вариант ускоренного универсального градиентного метода (Algorithm 7) с линейным поиском для задач с линейными ограничениями, доказал теорему о его сходимости (Theorem 4.1)).

**Публикации стандартного уровня:**

1. Chernov, A., Dvurechensky, P., and Gasnikov, A. Fast primal-dual gradient method for strongly convex minimization problems with linear constraints. In Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19-23, 2016, Proceedings (2016), Y. Kochetov, M. Khachay, V. Beresnev, E. Nurminski, and P. Pardalos, Eds., Springer International Publishing, pp. 391–403, Web of Science and Scopus (главный соавтор; автор диссертации предложил основной алгоритм и доказал теорему о его скорости сходимости).

**Доклады на конференциях и семинарах:**

1. International Workshop Advances in Optimization and Statistics, Берлин, 15.05.2014–16.05.2014, "Stochastic Intermediate Gradient Method for Convex Problems with Inexact Stochastic Oracle".

2. Семинар Modern Methods in Applied Stochastics and Nonparametric Statistics, Берлин, 03.06.2014, "Gradient methods for convex problems with stochastic inexact oracle".

3. V International Conference on Optimization Methods and Applications (OPTIMA–2014), Петровац, Черногория, 28.09.2014–04.10.2014, "Gradient-free optimization methods with ball randomization".

4. VI traditional school for young scientists Control, information, optimization, Москва, 22.06.2014–29.06.2014, "Gradient methods for convex problems with stochastic inexact oracle".

5. 38-th conference-school of IITP RAS Information technologies and systems, Нижний Новгород, 01.09.2014–05.09.2014, "Stochastic Intermediate Gradient Method for Convex Problems with Inexact Stochastic Oracle".

6. Workshop Frontiers of High Dimensional Statistics, Optimization, and Econometrics, Москва, 26.02.2015–27.02.2015, "Random gradient-free methods for random walk based web page ranking functions learning".

7. VII traditional school for young scientists Control, information, optimization, Москва, 14.06.2014-20.06.2014, "Semi-Supervised PageRank Model Learning with Gradient-Free Optimization Methods".

8. 29-th conference-school of IITP RAS Information technologies and systems, Сочи, 07.09.2014–11.09.2015, "Stochastic Intermediate Gradient Method: convex and strongly-convex case".

9. 30th annual conference of Belgian Operational Research Society (ORBEL 30), Бельгия, 28.01.2016–29.01.2016, "Random gradient-free methods for ranking algorithm learning".

10. Workshop on Modern Statistics and Optimization, Москва, 23.02.2016–24.02.2016, "Gradient and gradient-free methods for pagerank algorithm learning".

11. VII International Conference Optimization and Applications (OPTIMA 2016), Петровац, Черногория, 25.09.2016–02.10.2016, "Accelerated Primal-Dual Gradient Method for Linearly Constrained Minimization Problems".

12. VIII Moscow International Conference on Operations Research (ORM 2016), Москва, 17.10.2016–22.10.2016, "Accelerated Primal-Dual Gradient Method for Composite Optimization with Unknown Smoothness Parameter"

13. **Conference on Neural Information Processing Systems (NIPS 2016)**, Барселона, 05.12.2016–10.12.2016, "Learning Supervised PageRank with Gradient-Based and Gradient-Free Optimization Methods".

14. Workshop Shape, Images and Optimization, Мюнстер, Германия, 28.02.2017–03.03.2017, "Gradient Method With Inexact Oracle for Composite Non-Convex Optimization".

15. Optimization and Statistical Learning, Лез Уш, Франция, 10.04.2017-14.04.2017, "Gradient Method With Inexact Oracle for Composite Non-Convex Optimization".

16. Foundations of Computational Mathematics, Барселона, 10.07.2017-19.07.2017, "Gradient Method With Inexact Oracle for Composite Non-Convex Optimization".

17. Co-Evolution of Nature and Society Modelling, Problems & Experience. Devoted to Academician Nikita Moiseev centenary (Moiseev-100), Москва, 07.11.2017–10.11.2017, "Adaptive Similar Triangles Method: a Stable Alternative to Sinkhorn's Algorithm for Regularized Optimal Transport".

18. 18th French-German-Italian Conference on Optimization, Germany, 25.09.2017–28.09.2017, Падерборн, Германия, "Gradient method with inexact oracle for composite non-convex optimization"

19. 3. International Matheon Conference on Compressed Sensing and its Applications, Берлин, 04.12.2017–08.12.2017, "Adaptive Similar Triangles Method: a Stable Alternative to Sinkhorn's Algorithm for Regularized Optimal Transport".

20. Games, Dynamics and Optimization (GDO2018), Вена, 13.03.2018–15.03.2018, "Primal-Dual Methods for Solving Infinite -Dimensional Games".

21. **International Conference on Machine Learning (ICML 2018)**, Стокгольм, 10.07.2018–15.07.2018, "Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm".

22. 23rd International Symposium on Mathematical Programming, Бордо, Франция, 01.07.2018–06.07.2018, "Computational Optimal Transport: Accelerated Gradient Descent vs Sinkhorn".

23. Grenoble Optimization Days 2018: Optimization algorithms and applications in statistical learning, Гренобль, Франция, 28.06.2018–29.06.2018, "Faster algorithms for (regularized) optimal transport".

24. Statistical Optimal Transport Conference, Москва, 24.07.2018–25.07.2018, "Computational Optimal Transport: Accelerated Gradient Descent vs Sinkhorn's Algorithm".

25. **Conference on Neural Information Processing Systems (NIPS 2018)**, Монреаль, Канада, 02.12.2018–08.12.2018, "Decentralize and randomize: Faster algorithm for Wasserstein barycenters".

26. Optimization and Statistical Learning, Лез Уш, Франция, 24.03.2019–29.03.2019, "Distributed optimization for Wasserstein barycenter".

27. **International Conference on Machine Learning (ICML 2019)**, Лонг Бич, США, 09.06.2019–15.06.2019, "On the Complexity of Approximating Wasserstein Barycenters".

28. International Conference on Continuous Optimization (ICCOPT 2019), Берлин, 03.08.2019-08.08.2019, "A Unifying Framework for Accelerated Randomized Optimization Methods".

29. Workshop on optimization and applications, Москва, 27.09.2019, "Accelerated Alternating Minimization for Optimal Transport".

30. Recent advances in mass transportation, Москва, 23.09.2019–27.09.2019, "On the complexity of optimal transport problems".

31. Workshop by the GAMM Activity Group on Computational and Mathematical Methods in Data Science, Берлин, 24.10.2019–25.10.2019, "On the complexity of optimal transport problems".

32. HSE-Yandex autumn school on generative models, Москва, 26.11.2019–29.11.2019, "Optimization methods for optimal transport".

33. Workshop on Mathematics of Deep Learning 2019, Берлин, Germany, 03.12.2019–05.12.2019, "On the complexity of optimal transport problems".

34. Workshop on PDE Constrained Optimization under Uncertainty and Mean Field Games, Берлин, 28.01.2020–30.01.2020, "Distributed optimization for Wasserstein barycenters".

## 2 Оптимизация с неточным оракулом

В этом разделе приводится краткое содержание основных результатов по алгоритмам и их скорости сходимости для задач с неточной информацией. Рассматриваются методы первого порядка, безградиентные методы и методы с проиводной по направлению.

## 2.1 Стохастический промежуточный градиентный метод для задач выпуклой стохастической оптимизации

Результаты этого раздела опубликованы в работах [21, 22].

Пусть $E$ конечномерное векторное пространство, а $E^*$ его сопряженное. Значение линейной функции $g \in E^*$ в точке $x \in E$ обозначается как $\langle g, x\rangle$. Пусть $\|\cdot\|$ некоторая норма на $E$. Через $\|\cdot\|_*$ обозначим сопряженную к ней, т.е.
$\|g\|_* = \sup_{y\in E}\{\langle g, y\rangle : \|y\|_E \leq 1\}$. Через $\partial f(x)$ обозначим субдифференциал функции $f(x)$ в точке $x$. В этом разделе мы рассматриваем задачу *композитной оптимизации* вида

$$\min_{x\in Q}\{\varphi(x) := f(x) + h(x)\}, \tag{1}$$

где $Q \subset E$ замкнутое выпуклое множество, $h(x)$ простая выпуклая функция, $f(x)$ выпуклая функция со *стохастическим неточным оракулом* [23]. Это означает, что для любого $x \in Q$, существует $f_{\delta,L}(x) \in \mathbb{R}$ и $g_{\delta,L}(x) \in E^*$ такие, что

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

а также, что вместо $(f_{\delta,L}(x), g_{\delta,L}(x))$ (будем называть эту пару $(\delta, L)$-оракул), в алгоритмах можно использовать только их стохастическую аппроксимацию $(F_{\delta,L}(x,\xi), G_{\delta,L}(x,\xi))$. Последнее означает, что для любой точки $x \in Q$, есть некоторая случайная величина $\xi$, чье вероятностное распределение сосредоточено намножестве $\Xi \subset \mathbb{R}$, и такая, что $\mathbb{E}_\xi F_{\delta,L}(x,\xi) = f_{\delta,L}(x)$, $\mathbb{E}_\xi G_{\delta,L}(x,\xi) = g_{\delta,L}(x)$ и $\mathbb{E}_\xi(\|G_{\delta,L}(x,\xi) - g_{\delta,L}(x)\|_*)^2 \leq \sigma^2$.

В алгоритме используется *прокс-функция* $d(x)$, являющаяся дифференцируемой и сильно выпуклой с параметром 1 на $Q$ относительно нормы $\|\cdot\|$. Пусть $x_0$ точка минимума $d(x)$ на $Q$. Домножая на константу и смещая аргумент $d(x)$ если это необходимо, всегда можно добиться того, чтобы $d(x_0) = 0$, $d(x) \geq \frac{1}{2}\|x - x_0\|^2$, $\forall x \in Q$. Определим также соответствующую дивергенцию Брегмана: $V(x,z) = d(x) - d(z) - \langle \nabla d(z), x - z\rangle$. Пусть $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0} \subset \mathbb{R}$ три последовательности чисел, удовлетворяющие соотношениям

$$\alpha_0 \in ]0,1], \quad \beta_{i+1} \geq \beta_i > L, \quad \forall i \geq 0, \tag{3}$$

$$0 \leq \alpha_i \leq B_i, \quad \forall i \geq 0, \tag{4}$$

$$\alpha_k^2\beta_k \leq B_k\beta_{k-1} \leq \left(\sum_{i=0}^{k}\alpha_i\right)\beta_{k-1}, \quad \forall k \geq 1. \tag{5}$$

$$A_k := \sum_{i=0}^{k}\alpha_i, \quad \tau_i := \frac{\alpha_{i+1}}{B_{i+1}} \tag{6}$$

Стохастический промежуточный градиентный метод (СПГМ) описан ниже как Алгоритм 1. Пусть $a \geq 1$ и $b \geq 0$ некоторые параметры. Предположим, что известно число $R$ такое, что $\sqrt{2d(x^*)} \leq R$. Для $p \in [1, 2]$ положим

$$\alpha_i = \frac{1}{a}\left(\frac{i+p}{p}\right)^{p-1}, \quad \forall i \geq 0, \tag{7}$$

$$\beta_i = L + \frac{b\sigma}{R}(i+p+1)^{\frac{2p-1}{2}}, \quad \forall i \geq 0, \tag{8}$$

$$B_i = a\alpha_i^2 = \frac{1}{a}\left(\frac{i+p}{p}\right)^{2p-2}, \quad \forall i \geq 0. \tag{9}$$

**Теорема 2.1.** *Если последовательности* $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ *выбраны в соответствии с* (7), (8), (9) *с* $a = 2^{\frac{2p-1}{2}}$ *и* $b = 2^{\frac{5-2p}{4}} p^{\frac{1-2p}{2}}$, *тогда последовательность* $y_k$, *генерируемая СПГМ, удовлетворяет неравенствам*

$$\mathbb{E}_{\xi_0,\ldots,\xi_k}\varphi(y_k) - \varphi^* \leq \frac{LR^2 p^p 2^{\frac{2p-3}{2}}}{(k+p)^p} + \frac{\sigma R 2^{\frac{3+2p}{4}}\sqrt{p}(k+p+2)^{p-\frac{1}{2}}}{(k+p)^p} +$$
$$+ 2^{2p-1}\left(\left(\frac{k+p}{p}\right)^{p-1} + 1\right)\delta \leq \frac{C_1 LR^2}{k^p} + \frac{C_2\sigma R}{\sqrt{k}} + C_3 k^{p-1}\delta =$$
$$= \Theta\left(\frac{LR^2}{k^p} + \frac{\sigma R}{\sqrt{k}} + k^{p-1}\delta\right),$$

*где* $C_1 = 4\sqrt{2}$, $C_2 = 16\sqrt{2}$, $C_3 = 48$.

Получим также верхнюю оценку на вероятность больших уклонений для $\varphi(y_k) - \varphi^*$. Для этого сделаем следующие предположения.

1. $\xi_0, \ldots, \xi_k$ независимые, одинаково распределенные случайные величины.

2. $G_{\delta,L}(x, \xi)$ удовлетворяет субгауссовскому предположению

$$\mathbb{E}_\xi\left[\exp\left(\frac{\|G_{\delta,L}(x,\xi) - g_{\delta,L}(x)\|_*^2}{\sigma^2}\right)\right] \leq \exp(1).$$

3. Множество $Q$ ограничено и известно число $D > 0$ такое, что $\max_{x,y\in Q}\|x - y\| \leq D$.

**Теорема 2.2.** *Если последовательности* $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ *выбраны в соответствии с* (7), (8), (9) *с* $a = 2^{\frac{2p-1}{2}}$ *и* $b = 2^{\frac{5-2p}{4}} p^{\frac{1-2p}{2}}$, *тогда*

**Algorithm 1** Стохастический промежуточный градиентный метод (СПГМ)

**Вход:** Последовательности $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$, функции $d(x)$, $V(x,z)$.

**Выход:** Точка $y_k$.

1: Найти $x_0 := \arg\min_{x\in Q}\{d(x)\}$. Пусть $\xi_0$ реализация случайной величины $\xi$. Найти $G_{\delta,L}(x_0,\xi_0)$. Положить $k=0$.
2: $y_0 := \arg\min_{x\in Q}\{\beta_0 d(x) + \alpha_0\langle G_{\delta,L}(x_0,\xi_0), x - x_0\rangle + \alpha_0 h(x)\}$.
3: **repeat**
4: $z_k := \arg\min_{x\in Q}\{\beta_k d(x) + \sum_{i=0}^{k}\alpha_i\langle G_{\delta,L}(x_i,\xi_i), x - x_i\rangle + A_k h(x)\}$.
5: $x_{k+1} := \tau_k z_k + (1-\tau_k)y_k$.
6: Пусть $\xi_{k+1}$ реализация случайной величины $\xi$. Вычислить $G_{\delta,L}(x_{k+1},\xi_{k+1})$.
7: $\hat{x}_{k+1} := \arg\min_{x\in Q}\{\beta_k V(x,z_k) + \alpha_{k+1}\langle G_{\delta,L}(x_{k+1},\xi_{k+1}), x - z_k\rangle + \alpha_{k+1}h(x).\}$.
8: $w_{k+1} := \tau_k\hat{x}_{k+1} + (1-\tau_k)y_k$.
9: $y_{k+1} := \frac{A_{k+1}-B_{k+1}}{A_{k+1}}y_k + \frac{B_{k+1}}{A_{k+1}}w_{k+1}$.
10: **until**

*последовательность $y_k$, сгенерированная СПГМ, удовлетворяет*

$$\mathsf{P}\left(\varphi(y_k) - \varphi^* > \frac{C_1 LR^2}{k^p} + \frac{C_2(1+\Omega)\sigma R}{\sqrt{k}} + C_3 k^{p-1}\delta + \frac{C_4 D\sigma\sqrt{\Omega}}{\sqrt{k}}\right)$$
$$\leq \mathsf{P}\left(\varphi(y_k) - \varphi^* > \frac{LR^2 p^p 2^{\frac{2p-3}{2}}}{(k+p)^p} + \frac{(1+\Omega)\sigma R 2^{\frac{3+2p}{4}}\sqrt{p}(k+p+2)^{p-\frac{1}{2}}}{(k+p)^p}\right.$$
$$\left. + 2^{2p-1}\left(\left(\frac{k+p}{p}\right)^{p-1} + 1\right)\delta + \frac{2D\sigma\sqrt{6\Omega p}}{\sqrt{k+p}}\right) \leq 3\exp(-\Omega),$$

*где $C_1 = 4\sqrt{2}$, $C_2 = 16\sqrt{2}$, $C_3 = 48$, $C_4 = 4\sqrt{3}$.*

Рассмотрим две модификации СПГМ для сильно выпуклых задач. Для первой модификации оценивается скорость сходимости в терминах матожидания невязки по функции, а для второй модификации ограничивается вероятность больших уклонений от этой скорости сходимости. Для получения этих результатов дополнительно предположим, что $E$ есть Евклидово пространство со скалярным произведением $\langle\cdot,\cdot\rangle$ и нормой $\|x\| := \sqrt{\langle x, Hx\rangle}$, где $H$ симметричная положительно определенная матрица. Без ограничения общности предположим, что $d(x)$ удовлетворяет условиям $0 = \arg\min_{x\in Q} d(x)$ и $d(0) = 0$. Также предположим, что функция $\varphi(x)$ является сильно выпуклой, т.е.

$$\frac{\mu}{2}\|x - y\|^2 \leq \varphi(y) - \varphi(x) - \langle g(x), y - x\rangle$$

для всех $x, y \in Q, g(x) \in \partial\varphi(x)$. Из этого следует, что

$$\varphi(x) - \varphi(x^*) \geq \frac{\mu}{2}\|x - x^*\|^2, \quad \forall x \in Q, \tag{10}$$

где $x^*$ – решение задачи (1). Также предположим, что $d(x)$ удовлетворяет следующему свойству. Если $x_0$ случайный вектор, удовлетворяющий $\mathbb{E}_{x_0}\|x - x_0\|^2 \leq R_0^2$ для некоторой фиксированной точки $x$ и числа $R_0$, тогда существует число $V > 0$ такое, что

$$\mathbb{E}_{x_0} d\left(\frac{x - x_0}{R_0}\right) \leq \frac{V^2}{2}. \tag{11}$$

**Algorithm 2** Стохастический промежуточный градиентный метод для сильно выпуклых задач

**Вход:** Функция $d(x)$, точка $u_0$, число $R_0$ такое, что $\|u_0 - x^*\| \leq R_0$, число $p \in [1, 2]$.

**Выход:** Точка $u_{k+1}$.

1: Положить $k = 0$.

2: Вычислить

$$N_k := \left\lceil \left(\frac{4\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{1}{p}} \right\rceil. \tag{12}$$

3: **repeat**

4: Вычислить

$$m_k := \max\left\{1, \left\lceil \frac{16\mathrm{e}^{k+2} C_2^2 \sigma^2 V^2}{\mu^2 R_0^2 N_k} \right\rceil \right\}, \tag{13}$$

$$R_k^2 := R_0^2 \mathrm{e}^{-k} + \frac{2^p \mathrm{e} C_3 \delta}{\mu(\mathrm{e} - 1)} \left(\frac{4\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}} \left(1 - \mathrm{e}^{-k}\right). \tag{14}$$

5: Запустить Алгоритм 1 с $x_0 = u_k$ и прокс-функцией $d\left(\frac{x - u_k}{R_k}\right)$ на $N_k$ шагов с использованием оракула $\tilde{G}_{\delta,L}^k(x) := \frac{1}{m_k}\sum_{i=1}^{m_k} G_{\delta,L}(x, \xi^i)$ (где $\xi^i$, $i = 1, ..., m_k$ выборка независимых реализаций $\xi$) на каждом шаге и с последовательностями $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ определенными в Теореме 2.1.

6: Положить $u_{k+1} = y_{N_k}$, $k = k + 1$.

7: **until**

**Теорема 2.3.** *После $k \geq 1$ внешних итераций Алгоритма 2 справедли-*

*вы неравенства*

$$\mathbb{E}\varphi(u_k) - \varphi^* \leq \frac{\mu R_0^2}{2}\mathrm{e}^{-k} + \frac{C_3\mathrm{e}2^{p-1}}{\mathrm{e}-1}\left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\delta, \tag{15}$$

$$\mathbb{E}\|u_k - x^*\|^2 \leq R_0^2\mathrm{e}^{-k} + \frac{C_3\mathrm{e}2^{p}}{\mu(\mathrm{e}-1)}\left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\delta. \tag{16}$$

*Как следствие, если ошибка оракула $\delta$ выбрана удовлетворяющей неравенству*

$$\delta \leq \frac{\varepsilon(\mathrm{e}-1)}{2^pC_3\mathrm{e}}\left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{1-p}{p}}, \tag{17}$$

*то достаточно не более* $N = \left\lceil \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right\rceil$ *внешних итераций и не более чем*

$$\left(1 + \left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{1}{p}}\right)\left(1 + \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right) + \frac{16\mathrm{e}^3C_2^2\sigma^2V^2}{\mu\varepsilon(\mathrm{e}-1)}$$

*вызовов оракула для того, чтобы гарантировать выполнение неравенства* $\mathbb{E}\varphi(u_N) - \varphi^* \leq \varepsilon$.

Чтобы получить оценку сложности в терминах вероятностей больших уклонений, дополнительно предположим, что прокс-функция имеет квадратичный рост с параметром $V^2$ по отношению к выбранной норме, т.е.

$$d(x) \leq \frac{V^2}{2}\|x\|^2, \quad \forall x \in \mathbb{R}^n. \tag{18}$$

Ниже приведена модификация Алгоритма 2 для этого случая и теорема об оценке вероятностей больших уклонений для невязки по функции в точке, генерируемой этим алгоритмом.

**Теорема 2.4.** *После $N$ внешних итераций Алгоритма 3 справедливо неравенство*

$$\mathsf{P}\left\{\varphi(u_N) - \varphi^* > \frac{\mu R_0^2}{2}\mathrm{e}^{-N} + \frac{2^{p-1}\mathrm{e}C_3\delta}{(\mathrm{e}-1)}\left(\frac{6\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\delta\right\} \leq \Lambda. \tag{23}$$

*Как следствие, если неточность оракула $\delta$ удовлетворяет неравенству*

$$\delta \leq \frac{\varepsilon(\mathrm{e}-1)}{2^pC_3\mathrm{e}}\left(\frac{6\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{1-p}{p}}, \tag{24}$$

**Algorithm 3** Стохастический промежуточный градиентный метод для сильно выпуклых задач 2

**Вход:** Функция $d(x)$, точка $u_0$, число $R_0$ такое, что $\|u_0 - x^*\| \leq R_0$, число $p \in [1, 2]$, число $N \geq 1$ внешних итераций, доверительный уровень $\Lambda$.

**Выход:** Точка $u_N$.

1: Положить $k = 0$.

2: Вычислить

$$N_k := \left\lceil \left( \frac{6\mathrm{e} C_1 L V^2}{\mu} \right)^{\frac{1}{p}} \right\rceil. \tag{19}$$

3:

4: **repeat**

5: Вычислить

$$m_k := \max \left\{ 1, \left\lceil \frac{36\mathrm{e}^{k+2} C_2^2 \sigma^2 V^2 \left(1 + \ln\left(\frac{3N}{\Lambda}\right)\right)^2}{\mu^2 R_0^2 N_k} \right\rceil, \left\lceil \frac{144\mathrm{e}^{k+2} C_4^2 \sigma^2 \ln\left(\frac{3N}{\Lambda}\right)}{\mu^2 R_0^2 N_k} \right\rceil \right\}, \tag{20}$$

$$R_k^2 := R_0^2 \mathrm{e}^{-k} + \frac{2^p \mathrm{e} C_3 \delta}{\mu(\mathrm{e}-1)} \left( \frac{6\mathrm{e} C_1 L V^2}{\mu} \right)^{\frac{p-1}{p}} \left(1 - \mathrm{e}^{-k}\right), \tag{21}$$

$$Q_k := \left\{ x \in Q : \|x - u_k\|^2 \leq R_k^2 \right\}. \tag{22}$$

6: Запустить Алгоритм 1 для задачи $\min_{x \in Q_k} \varphi(x)$ с $x_0 = u_k$ и прокс-функцией $d\left(\frac{x-u_k}{R_k}\right)$ на $N_k$ итераций с использованием оракула $\tilde{G}_{\delta,L}^k(x) := \frac{1}{m_k} \sum_{i=1}^{m_k} G_{\delta,L}(x, \xi^i)$ (где $\xi^i$, $i = 1, ..., m_k$ независимые реализации случайной величины $\xi$) на каждом шаге и последовательностями $\{\alpha_i\}_{i \geq 0}$, $\{\beta_i\}_{i \geq 0}$, $\{B_i\}_{i \geq 0}$ определенными в Теореме 2.1.

7: Положить $u_{k+1} = y_{N_k}$, $k = k + 1$.

8: **until** $k = N - 1$

*то достаточно не более* $N = \left\lceil \ln\left(\frac{\mu R_0^2}{\varepsilon}\right) \right\rceil$ *внешних итераций и не более*

$$\left(1 + \left(\frac{6\mathrm{e} C_1 L V^2}{\mu}\right)^{\frac{1}{p}}\right) \left(1 + \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right) + \\ + \frac{36\mathrm{e}^3 C_2^2 \sigma^2 V^2}{\mu(\mathrm{e}-1)\varepsilon} \left(1 + \ln\left(\frac{3}{\Lambda}\left(1 + \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right)\right)\right)^2 + \\ + \frac{144\mathrm{e}^3 C_4^2 \sigma^2}{\mu\varepsilon(\mathrm{e}-1)} \ln\left(\frac{3}{\Lambda}\left(1 + \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right)\right) \tag{25}$$

*вызовов оракула для того, чтобы гарантировать выполнение неравен-*

*ства* $\mathsf{P}\{\varphi(u_N) - \varphi^* > \varepsilon\} \leq \Lambda$.

## 2.2 Обучение модели ранжирования веб-страниц градиентными и безградиентными методами

В этом разделе рассматривается параметрическая модель ранжирования веб-страниц и обучение параметров этой модели по данным в рамках подхода обучения с учителем. Результаты опубликованы в статье [24].

### 2.2.1 Постановка задачи минимизации риска

Рассматривается минимизация потерь при обучении, заданных функцией

$$f(\varphi) = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \|(A_q \pi_q(\varphi))_+\|_2^2 \tag{26}$$

от параметра $\varphi \in \mathbb{R}^m$ на некотором допустимом множестве $\Phi$. Здесь вектор $x_+$ имеет компоненты $[x_+]_i = \max\{x_i, 0\}$, числа $q, r_q$ и матрицы $A_q \in \mathbb{R}^{r_q \times p_q}, q \in Q$ заданы. Обозначим $r = \max_{q \in Q} r_q$. Кроме того, векторы распределения вероятностей $\pi_q(\varphi) \in \mathbb{R}^{p_q}$ заданы как решение уравнения

$$\pi = \alpha \pi_q^0(\varphi) + (1 - \alpha) P_q^T(\varphi) \pi, \tag{27}$$

где $\pi_q^0(\varphi) \in \mathbb{R}^{p_q}$ – заданная дифференцируемая вектор-функция с первыми $n_q$ ненулевыми компонентами и остальными компонентами равными нулю, $P_q(\varphi) \in \mathbb{R}^{p_q \times p_q}$ – заданная дифференцируемая матричнозначная функция. Обозначим $p = \max_{q \in Q} p_q$, $n = \max_{q \in Q} n_q$, $s = \max_{q \in Q} s_q$, где $s_q$ – максимальное число ненулевых компонент в строках матрицы $P_q$.

Выберем некоторый вектор $\hat{\varphi}$ и число $R > 0$ такие, что множество $\Phi$ заданное как $\Phi = \{\varphi \in \mathbb{R}^m : \|\varphi - \hat{\varphi}\|_2 \leq R\}$ содержится во множестве $\mathbb{R}^m_{++}$ векторов с положительными компонентами. Таким образом, решается следующая задача минимизации потерь при обучении

$$\min_{\varphi \in \Phi} f(\varphi), \Phi = \{\varphi \in \mathbb{R}^m : \|\varphi - \hat{\varphi}\|_2 \leq R\}. \tag{28}$$

Алгоритм [25] для аппроксимации вектора $\pi_q(\varphi)$ при фиксированном $q \in Q$ строит последовательность $\pi_k$ и как результат генерирует вектор $\tilde{\pi}_q(\varphi, N)$, где $N$ фиксированное натуральное число, заданные как

$$\pi_0 = \pi_q^0(\varphi), \quad \pi_{k+1} = P_q^T(\varphi)\pi_k, \quad \tilde{\pi}_q(\varphi, N) = \frac{\alpha}{1 - (1 - \alpha)^{N+1}} \sum_{k=0}^{N} (1 - \alpha)^k \pi_k. \tag{29}$$

**Лемма 2.1.** *Предположим, что для некоторого $\delta_1 > 0$ алгоритм* (29) *с $N = \left\lceil \frac{1}{\alpha} \ln \frac{8r}{\delta_1} \right\rceil - 1$ используется для вычисления вектора $\tilde{\pi}_q(\varphi, N)$ для каждого $q \in Q$. Тогда*

$$\widetilde{f}(\varphi, \delta_1) = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \|(A_q \tilde{\pi}_q(\varphi, N))_+\|_2^2 \tag{30}$$

*satisfies*

$$|\widetilde{f}(\varphi, \delta_1) - f(\varphi)| \le \delta_1. \tag{31}$$

*Кроме того, вычисление $\widetilde{f}(\varphi, \delta_1)$ требует не более, чем $|Q|(3mps + 3psN + 6r)$ арифметических операций (а.о.) и не более $3ps$ ячеек памяти.*

Предлагаемое обобщение метода из [25] для вычисления $\frac{d\pi_q(\varphi)}{d\varphi^T}$ при любом $q \in Q$ состоит в следующем. Выбирается натуральное число $N_1$ и вычисляется $\tilde{\pi}_q(\varphi, N_1)$ согласно (29). Далее вычисляется последовательность $\Pi_k$ согласно

$$\Pi_0 = \alpha \frac{d\pi_q^0(\varphi)}{d\varphi^T} + (1-\alpha) \sum_{i=1}^{p_q} \frac{dp_i(\varphi)}{d\varphi^T} [\tilde{\pi}_q(\varphi, N_1)]_i, \quad \Pi_{k+1} = P_q^T(\varphi) \Pi_k. \tag{32}$$

Выходом алгоритма является (при некотором заданном натуральном $N_2$)

$$\tilde{\Pi}_q(\varphi, N_2) = \frac{1}{1 - (1-\alpha)^{N_2+1}} \sum_{k=0}^{N_2} (1-\alpha)^k \Pi_k. \tag{33}$$

Ниже используется следующая норма на пространстве матриц $A \in \mathbb{R}^{n_1 \times n_2}$: $\|A\|_1 = \max_{j=1,\dots,n_2} \sum_{i=1}^{n_1} |a_{ij}|$.

**Лемма 2.2.** *Пусть $\beta_1$ есть число (его можно вычислить явно, см. [24]) такое, что для всех $\varphi \in \Phi$ справедливо*

$$\alpha \left\| \frac{d\pi_q^0(\varphi)}{d\varphi^T} \right\|_1 + (1-\alpha) \sum_{i=1}^{p_q} \left\| \frac{dp_i(\varphi)}{d\varphi^T} \right\|_1 \le \beta_1. \tag{34}$$

*Предположим, что метод* (29) *с $N_1 = \left\lceil \frac{1}{\alpha} \ln \frac{24\beta_1 r}{\alpha \delta_2} \right\rceil - 1$ используется для всех $q \in Q$, чтобы вычислить вектор $\tilde{\pi}_q(\varphi, N_1)$ и метод* (32), (33) *с $N_2 = \left\lceil \frac{1}{\alpha} \ln \frac{8\beta_1 r}{\alpha \delta_2} \right\rceil - 1$ используется для всех $q \in Q$ для вычисления матриц $\tilde{\Pi}_q(\varphi, N_2)$* (33). *Тогда вектор*

$$\tilde{g}(\varphi, \delta_2) = \frac{2}{|Q|} \sum_{q=1}^{|Q|} \left( \tilde{\Pi}_q(\varphi, N_2) \right)^T A_q^T (A_q \tilde{\pi}_q(\varphi, N_1))_+ \tag{35}$$

*удовлетворяет*

$$\|\tilde{g}(\varphi, \delta_2) - \nabla f(\varphi)\|_\infty \leq \delta_2. \tag{36}$$

*Кроме того, вычисление $\tilde{g}(\varphi, \delta_2)$ требует не более, чем $|Q|(10mps + 3psN_1 + 3mpsN_2 + 7r)$ арифметических операций и не более $4ps + 4mp + r$ ячеек памяти.*

Таким образом, для рассматриваемой задачи минимизации функции потерь при обучении доступен неточный оракул. Ниже рассматриваются сначала общие алгоритмы для задач с неточным оракулом нулевого и первого порядка, а затем эти алгоритмы применяются для решения задачи обучения.

### 2.2.2 Решение задачи обучения методами нулевого порядка

В этом разделе сначала рассматривается общий метод нулевого порядка с неточным вычислением значения функции, а затем этот метод применяется для решения описанной выше задачи обучения. Пусть $\mathcal{E}$ – $m$-мерное векторное пространство. Рассмотрим функцию $f(\cdot) : \mathcal{E} \to \mathbb{R}$ и обозначим ее аргумент $x$ или $y$. Значение линейной функции $g \in \mathcal{E}^*$ в точке $x \in \mathcal{E}$ обозначается $\langle g, x \rangle$. Пусть выбрана некоторая норма $\|\cdot\|$ на $\mathcal{E}$. Будем говорить, что $f \in C_L^{1,1}(\|\cdot\|)$, если

$$|f(x) - f(y) - \langle \nabla f(y), x - y \rangle| \leq \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in \mathcal{E}. \tag{37}$$

Рассматривается задача $\min_{x \in X} f(x)$, где $f \in C_L^{1,1}(\|\cdot\|)$, $X$ – замкнутое выпуклое множество и существует число $D \in (0, +\infty)$ такое, что $\mathrm{diam}X := \max_{x,y \in X} \|x - y\| \leq D$. Также предположим, что неточный оракул нулевого порядка для $f(x)$ в любой точке $x \in X$ выдает значение $\widetilde{f}(x, \delta) = f(x) + \tilde{\delta}(x)$, где $\tilde{\delta}(x)$ – ошибка, удовлетворяющая для некоторого известного $\delta > 0$ неравенству $|\tilde{\delta}(x)| \leq \delta$ для всех $x \in X$. Обозначим $x^* \in \arg\min_{x \in X} f(x)$ и $f^* = \min_{x \in X} f(x)$.

В отличие от [17], определим смещенную безградиентную аппроксимацию градиента $g_\tau(x, \delta) = \frac{m}{\tau}(\widetilde{f}(x + \tau\xi, \delta) - \widetilde{f}(x, \delta))\xi$, где $\xi$ – случайный вектор равномерно распределенный на единичной Евклидовой сфере $\mathcal{S} = \{t \in \mathbb{R}^m : \|t\|_2 = 1\}$, $\tau$ – параметр сглаживания.

Алгоритм 4 ниже является модификацией метода проекции градиента. $\Pi_X(x)$ обозначает Евклидову проекцию точки $x$ на множество $X$.

Следующая теорема является результатом о скорости сходимости Алгоритма 4. Обозначим через $\Xi_k = (\xi_0, \ldots, \xi_k)$ историю реализаций вектора $\xi$, генерируемых на каждой итерации алгоритма.

**Теорема 2.5.** *Пусть $f \in C_L^{1,1}(\|\cdot\|_2)$ и выпукла. Предположим, что $x^* \in \mathrm{int}X$, и последовательность $x_k$ генерируется Алгоритмом 4 с $h = \frac{1}{8mL}$.*

**Algorithm 4** Метод градиентного типа

**Вход:** Точка $x_0 \in X$, шаг $h > 0$, число шагов $M$.
Положить $k = 0$.
**repeat**
  Сгенерировать $\xi_k$ и вычислить соответствующий вектор $g_\tau(x_k, \delta)$.
  Вычислить $x_{k+1} = \Pi_X(x_k - hg_\tau(x_k, \delta))$.
  Положить $k = k + 1$.
**until** $k > M$
**Выход:** Точка $y_M = \arg\min_x\{f(x) : x \in \{x_0, \dots, x_M\}\}$.

*Тогда для любого $M \geq 0$ справедливо неравенство*

$$\mathbb{E}_{\Xi_{M-1}} f(y_M) - f^* \leq \frac{8mLD^2}{M+1} + \frac{\tau^2 L(m+8)}{8} + \frac{\delta m D}{4\tau} + \frac{\delta^2 m}{L\tau^2}. \tag{38}$$

**Algorithm 5** Безградиентный метод для задачи (28)

**Вход:** Точка $\varphi_0 \in \Phi$, $L$ – константа Липшица градиента $f(\varphi)$ на $\Phi$, точность $\varepsilon > 0$.
Положить $M = \left\lceil 128m\frac{LR^2}{\varepsilon} \right\rceil$, $\delta = \frac{\varepsilon^{\frac{3}{2}}\sqrt{2}}{16mR\sqrt{L(m+8)}}$, $\tau = \sqrt{\frac{2\varepsilon}{L(m+8)}}$.
Положить $k = 0$.
**repeat**
  Сгенерировать случайный вектор $\xi_k$ равномерно распределенный на единичной Евклидовой сфере $\mathcal{S}$ в $R^m$.
  Вычислить $\widetilde{f}(\varphi_k + \tau\xi_k, \delta)$, $\widetilde{f}(\varphi_k, \delta)$, используя Лемму 2.1 с $\delta_1 = \delta$.
  Вычислить $g_\tau(\varphi_k, \delta) = \frac{m}{\tau}(\widetilde{f}(\varphi_k + \tau\xi_k, \delta) - \widetilde{f}(\varphi_k, \delta))\xi_k$.
  Вычислить $\varphi_{k+1} = \Pi_\Phi\left(\varphi_k - \frac{1}{8mL} g_\tau(\varphi_k, \delta)\right)$.
  Положить $k = k + 1$.
**until** $k > M$
**Выход:** Точка $\hat{\varphi}_M = \arg\min_\varphi\{f(\varphi) : \varphi \in \{\varphi_0, \dots, \varphi_M\}\}$.

Применим общий безградиентный метод для решения задачи обучения (28). Получившийся алгоритм приведен как Алгоритм 5.

**Теорема 2.6.** *Предположим, что множество $\Phi$ в (28) выбрано так, что $f(\varphi)$ выпукла на $\Phi$ и некоторое решение $\varphi^* \in \operatorname{Arg\,min}_{\varphi\in\Phi} f(\varphi)$ принадлежит также* $\operatorname{int}\Phi$. *Тогда среднее общее число арифметических операций в Алгоритме 5 при заданной точности $\varepsilon$ (т.е. для того, чтобы выполнилось неравенство $\mathbb{E}_{\Xi_{M-1}} f(\hat{\varphi}_M) - f(\varphi^*) \leq \varepsilon$) не превосходит*

$$768mps|Q|\frac{LR^2}{\varepsilon}\left(m + \frac{1}{\alpha}\ln\frac{128mrR\sqrt{L(m+8)}}{\varepsilon^{3/2}\sqrt{2}} + 6r\right).$$

### 2.2.3 Решение задачи обучения методом первого порядка

В этом разделе сначала рассматривается общий метод первого порядка с неточными значениями градиента, а затем этот метод применяется к решению задачи обучения. Пусть $\mathcal{E}$ – конечномерное векторное пространство и $\mathcal{E}^*$ – его сопряженное. Значение линейной функции $g \in \mathcal{E}^*$ в точке $x \in \mathcal{E}$ обозначается как $\langle g, x \rangle$. Пусть $\|\cdot\|$ – некоторая норма на $\mathcal{E}$, а $\|\cdot\|_*$ – ее сопряженная. Целью является решение следующей задачи *композитной оптимизации*

$$\min_{x \in X} \{\psi(x) := f(x) + h(x)\}, \tag{39}$$

где $X \subset \mathcal{E}$ – замкнутое выпуклое множество, $h(x)$ – простая выпуклая функция, например, $\|x\|_1$. Предполагается, что функция $f(x)$ доступна через неточный оракул первого порядка в следующем смысле. Существует число $L \in (0, +\infty)$ такое, что для любого $\delta \geq 0$ и любого $x \in X$ доступны $\widetilde{f}(x, \delta) \in \mathbb{R}$ и $\tilde{g}(x, \delta) \in \mathcal{E}^*$ удовлетворяющие

$$|f(y) - (\widetilde{f}(x, \delta) - \langle \tilde{g}(x, \delta), y - x \rangle)| \leq \frac{L}{2}\|x - y\|^2 + \delta. \tag{40}$$

для любого $y \in X$. Константа $L$ рассматривается как обобщение константы Липшица градиента $f$ так как в случае точного оракула первого порядка для $f \in C^{1,1}_L(\|\cdot\|)$ неравенство (40) выполнено с $\delta = 0$. Описанный неточный оракул является обобщением концепции $(\delta, L)$-оракула, предложенной в [26] для выпуклых задач.

Выберем прокс-функцию $d(x)$, являющуюся непрерывно дифференцируемой и 1-сильно выпуклой на $X$ в норме $\|\cdot\|$, т.е. для любых $x, y \in X$ $d(y) - d(x) - \langle \nabla d(x), y - x \rangle \geq \frac{1}{2}\|y - x\|^2$. Также определим соответствующую дивергенцию Брегмана $V(x, z) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$.

**Теорема 2.7.** *Предположим, что для $f(x)$ доступен неточный оракул первого порядка в смысле* (40)*, а также, что существует число $\psi^* > -\infty$ такое, что $\psi(x) \geq \psi^*$ для всех $x \in X$. Тогда после $M$ итераций Алгоритма 6 выполнено неравенство*

$$\|M_K(x_K - x_{K+1})\|^2 \leq \frac{4L(\psi(x_0) - \psi^*)}{M + 1} + \frac{\varepsilon}{2}. \tag{41}$$

*Кроме того, общее число внутренних итераций не превышает $M + \log_2 \frac{2L}{L_0}$.*

Применим описанный выше общий метод к решению задачи обучения. Положим $\mathcal{E} = R^m$ и $\|\cdot\| = \|\cdot\|_2$, выберем $d(\varphi) = \frac{1}{2}\|\varphi\|_2^2$ и $V(\varphi, \omega) = \frac{1}{2}\|\varphi - \omega\|_2^2$. Получившийся алгоритм приведен как Алгоритм 7.

**Algorithm 6** Адаптивный метод проекции градиента

**Вход:** Точка $x_0 \in X$, число $L_0 > 0$.
Положить $k = 0$, $z = +\infty$.
**repeat**
  Положить $M_k = L_k$, flag $= 0$.
  **repeat**
    Положить $\delta = \frac{\varepsilon}{16 M_k}$. Вычислить $\widetilde{f}(x_k, \delta)$ и $\tilde{g}(x_k, \delta)$.
    $w_k = \arg\min_{x \in Q} \{\langle \tilde{g}(x_k, \delta), x\rangle + M_k V(x, x_k) + h(x)\}$
    Если неравенство

$$\widetilde{f}(w_k, \delta) \leq \widetilde{f}(x_k, \delta) + \langle \tilde{g}(x_k, \delta), w_k - x_k\rangle + \frac{M_k}{2}\|w_k - x_k\|^2 + \frac{\varepsilon}{8 M_k}$$

    выполнено, то flag $= 1$. Иначе $M_k = 2M_k$.
  **until** flag $= 1$
  Положить $x_{k+1} = w_k$, $L_{k+1} = \frac{M_k}{2}$.
  Если $\|M_k(x_k - x_{k+1})\| < z$, то $z = \|M_k(x_k - x_{k+1})\|$, $K = k$.
  Положить $k = k + 1$.
**until** $z \leq \varepsilon$
**Выход:** Точка $x_{K+1}$.

**Теорема 2.8.** *Общее число арифметических операций в Алгоритме 7 для достижения точности $\varepsilon$ (т.е. для того, чтобы выполнилось неравенство*
$\|M_K(\varphi_K - \varphi_{K+1})\|_2^2 \leq \varepsilon$*) не превышает*

$$\left(\frac{8L(f(\varphi_0) - f^*)}{\varepsilon} + \log_2 \frac{2L}{L_0}\right) \cdot \left(7r|Q| + \frac{6mps|Q|}{\alpha} \ln \frac{1024\beta_1 r R L\sqrt{m}}{\alpha\varepsilon}\right).$$

## 2.3 Ускоренный рандомизированный спуск по направлению для задач выпуклой гладкой стохастической оптимизации

В этом разделе рассматриваются спуски по направлению с неточным оракулом для гладкой стохастической оптимизации. Полученные результаты опубликованы в статье [27]. Рассматривается следующая задача оптимизации

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) := \mathbb{E}_\xi[F(x, \xi)] = \int_{\mathcal{X}} F(x, \xi) dP(x) \right\}, \tag{42}$$

где $\xi$ – случайный вектор с вероятностным распределением $P(\xi)$, $\xi \in \mathcal{X}$, и почти наверное относительно $P$ функция $F(x, \xi)$ является замкнутой и выпуклой. Также предполагается, что почти наверное относительно

**Algorithm 7** Адаптивный градиентный метод для задачи (28)

**Вход:** Точка $\varphi_0 \in \Phi$, число $L_0 > 0$, точность $\varepsilon > 0$.
Положить $k = 0$, $z = +\infty$.
**repeat**
  Положить $M_k = L_k$, flag $= 0$.
  **repeat**
    Положить $\delta_1 = \frac{\varepsilon}{32M_k}$, $\delta_2 = \frac{\varepsilon}{64M_k R\sqrt{m}}$.
    Вычислить $\widetilde{f}(\varphi_k, \delta_1)$, используя Лемму 2.1, и $\tilde{g}(\varphi_k, \delta_2)$, используя Лемму 2.2.
    Найти

$$\omega_k = \arg\min_{\varphi \in \Phi} \left\{ \langle \tilde{g}(\varphi_k, \delta_2), \varphi \rangle + \frac{M_k}{2} \|\varphi - \varphi_k\|_2^2. \right\}$$

    Вычислить $\widetilde{f}(\omega_k, \delta_1)$, используя Лемму 2.1.
    Если неравенство

$$\widetilde{f}(\omega_k, \delta_1) \leq \widetilde{f}(\varphi_k, \delta_1) + \langle \tilde{g}(\varphi_k, \delta_2), \omega_k - \varphi_k \rangle + \frac{M_k}{2} \|\omega_k - \varphi_k\|_2^2 + \frac{\varepsilon}{8M_k}$$

    выполнено, то flag $= 1$. Иначе $M_k = 2M_k$.
  **until** flag $= 1$
  Положить $\varphi_{k+1} = \omega_k$, $L_{k+1} = \frac{M_k}{2}$, .
  Если $\|M_k(\varphi_k - \varphi_{k+1})\|_2 < z$, то $z = \|M_k(\varphi_k - \varphi_{k+1})\|_2$, $K = k$.
  Положить $k = k + 1$.
**until** $z \leq \varepsilon$
**Выход:** Точка $\varphi_{K+1}$.

$P$ функция $F(x, \xi)$ обладает градиентом $g(x, \xi)$, который является $L(\xi)$-непрерывным по Липшицу в Евклидовой норме, и что существует такое число $L_2 \geqslant 0$, что $\sqrt{\mathbb{E}_\xi L(\xi)^2} \leqslant L_2 < +\infty$. При этих предположениях $\mathbb{E}_\xi g(x, \xi) = \nabla f(x)$ и $f$ имеет $L_2$-Липшицев градиент по отношению к Евклидовой норме. Также предположим, что

$$\mathbb{E}_\xi[\|g(x, \xi) - \nabla f(x)\|_2^2] \leqslant \sigma^2, \tag{43}$$

где $\|\cdot\|_2$ – Евклидова норма.

Также предполагается, что оптимизационная процедура для заданной точки $x \in \mathbb{R}^n$, направления $e \in S_2(1)$ и независимой реализации $\xi$ может получить зашумленную стохастическую аппроксимацию $\widetilde{f}'(x, \xi, e)$

для производной по направлению $\langle g(x,\xi), e\rangle$:

$$\begin{aligned} \widetilde{f}'(x,\xi,e) &= \langle g(x,\xi), e\rangle + \zeta(x,\xi,e) + \eta(x,\xi,e), \\ \mathbb{E}_\xi(\zeta(x,\xi,e))^2 &\leqslant \Delta_\zeta,\ \forall x \in \mathbb{R}^n, \forall e \in S_2(1), \\ |\eta(x,\xi,e)| &\leqslant \Delta_\eta,\ \forall x \in \mathbb{R}^n, \forall e \in S_2(1),\ \text{п.н. по } \xi, \end{aligned} \tag{44}$$

где $S_2(1)$ – Евклидова сфера радиуса 1 с центром в нуле и величины ошибки $\Delta_\zeta$, $\Delta_\eta$ контролируются и могут быть сделаны сколь угодно маленькими. Отметим, что мы используем гладкость функции $F(\cdot,\xi)$ для того, чтобы записать производную по направлению как $\langle g(x,\xi), e\rangle$, но при этом доступ к стохастическому градиенту $g(x,\xi)$ не предполагается. Выберем прокс-функцию $d(x)$, являющуюся непрерывной, выпуклой на $\mathbb{R}^n$ и 1-сильно выпуклой на $\mathbb{R}^n$ в норме $\|\cdot\|_p$, т.е. такой, что для любых $x, y \in \mathbb{R}^n$ справедливо $d(y) - d(x) - \langle \nabla d(x), y - x\rangle \geq \frac{1}{2}\|y - x\|_p^2$. Без ограничения общности предполагается, что $\min\limits_{x\in\mathbb{R}^n} d(x) = 0$. Также определим соответствующую дивергенцию Брегмана $V[z](x) = d(x) - d(z) - \langle \nabla d(z), x - z\rangle$, $x, z \in \mathbb{R}^n$. Для случая $p = 1$ используется следующая прокс-функция [28]

$$d(x) = \frac{en^{(\kappa-1)(2-\kappa)/\kappa}\ln n}{2}\|x\|_\kappa^2, \quad \kappa = 1 + \frac{1}{\ln n}, \tag{45}$$

а для случая $p = 2$ используется $d(x) = \frac{1}{2}\|x\|_2^2$.

На основе стохастических реализаций (44) производной по направлению формируется следующая стохастическая аппроксимация $\nabla f(x)$

$$\widetilde{\nabla}^m f(x) = \frac{1}{m}\sum_{i=1}^m \widetilde{f}'(x,\xi_i,e)e, \tag{46}$$

где $e \in RS_2(1)$, $\xi_i$, $i = 1, ..., m$ независимые реализации $\xi$, $m$ – размер батча.

### 2.3.1 Алгоритмы и основные результаты для выпуклых задач

Ускоренный рандомизированный спуск по направлению (УРСН) приведен как Алгоритм 8.

**Теорема 2.9.** *Пусть УРСН применен к задаче* (42). *Тогда*

$$\begin{aligned} \mathbb{E}[f(y_N)] - f(x^*) \leqslant{}& \frac{384\Theta_p n^2 \rho_n L_2}{N^2} + \frac{4N}{nL_2}\cdot\frac{\sigma^2}{m} + \frac{61N}{24L_2}\Delta_\zeta + \frac{122N}{3L_2}\Delta_\eta^2 \\ &+ \frac{12\sqrt{2n\Theta_p}}{N^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2, \end{aligned} \tag{47}$$

*где* $\Theta_p = V[z_0](x^*)$ *определено выбранной прокс-функцией и* $\mathbb{E}[\cdot] = \mathbb{E}_{e_1,\ldots,e_N,\xi_{1,1},\ldots,\xi_{N,m}}[\cdot]$.

**Algorithm 8** Ускоренный рандомизированный спуск по направлению (УРСН)

**Вход:** $x_0$ —начальная точка; $N \geqslant 1$ — число итераций; $m \geqslant 1$ — размер батча.

**Выход:** Точка $y_N$.

$y_0 \leftarrow x_0$, $z_0 \leftarrow x_0$.
**for** $k = 0, \ldots, N-1$. **do**
  $\alpha_{k+1} \leftarrow \frac{k+2}{96n^2\rho_n L_2}$, $\tau_k \leftarrow \frac{1}{48\alpha_{k+1}n^2\rho_n L_2} = \frac{2}{k+2}$.
  Сгенерировать $e_{k+1} \in RS_2(1)$ независимо от предыдущих итераций и $\xi_i$, $i = 1, ..., m$ – независимые реализации $\xi$.
  $\widetilde{\nabla}^m f(x_{k+1}) = \frac{1}{m}\sum\limits_{i=1}^{m} \widetilde{f}'(x_{k+1}, \xi_i, e)e$.
  $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k)y_k$.
  $y_{k+1} \leftarrow x_{k+1} - \frac{1}{2L_2}\widetilde{\nabla}^m f(x_{k+1})$.
  $z_{k+1} \leftarrow \underset{z\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\alpha_{k+1} n \left\langle \widetilde{\nabla}^m f(x_{k+1}),\, z - z_k \right\rangle + V[z_k](z)\right\}$.
**end for**
**return** $y_N$

**Algorithm 9** Рандомизированный спуск по направлению (РСН)

**Вход:** $x_0$ —начальная точка; $N \geqslant 1$ — число итераций; $m \geqslant 1$ — размер батча.

**Выход:** Точка $\bar{x}_N$.

**for** $k = 0, \ldots, N-1$. **do**
  $\alpha \leftarrow \frac{1}{48n\rho_n L_2}$.
  Сгенерировать $e_{k+1} \in RS_2(1)$ независимо от предыдущих итераций и $\xi_i$, $i = 1, ..., m$ – независимые реализации $\xi$.
  $\widetilde{\nabla}^m f(x_k) = \frac{1}{m}\sum\limits_{i=1}^{m} \widetilde{f}'(x_k, \xi_i, e)e$.
  $x_{k+1} \leftarrow \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\alpha n \left\langle \widetilde{\nabla}^m f(x_k),\, x - x_k \right\rangle + V[x_k](x)\right\}$.
**end for**
**return** $\bar{x}_N \leftarrow \frac{1}{N}\sum\limits_{k=0}^{N-1} x_k$

| | $p=1$ | $p=2$ |
|---|---|---|
| $N$ | $\sqrt{\frac{n\ln nL_2\Theta_1}{\varepsilon}}$ | $\sqrt{\frac{n^2L_2\Theta_2}{\varepsilon}}$ |
| $m$ | $\max\left\{1, \sqrt{\frac{\ln n}{n}}\cdot\frac{\sigma^2}{\varepsilon^{3/2}}\cdot\sqrt{\frac{\Theta_1}{L_2}}\right\}$ | $\max\left\{1, \frac{\sigma^2}{\varepsilon^{3/2}}\cdot\sqrt{\frac{\Theta_2}{L_2}}\right\}$ |
| $\Delta_\zeta$ | $\min\left\{n(\ln n)^2L_2^2\Theta_1,\ \frac{\varepsilon^2}{n\Theta_1},\ \frac{\varepsilon^{\frac{3}{2}}}{\sqrt{n\ln n}}\cdot\sqrt{\frac{L_2}{\Theta_1}}\right\}$ | $\min\left\{n^3L_2^2\Theta_2,\ \frac{\varepsilon^2}{n\Theta_2},\ \frac{\varepsilon^{\frac{3}{2}}}{n}\cdot\sqrt{\frac{L_2}{\Theta_2}}\right\}$ |
| $\Delta_\eta$ | $\min\left\{\sqrt{n}\ln nL_2\sqrt{\Theta_1},\ \frac{\varepsilon}{\sqrt{n\Theta_1}},\ \frac{\varepsilon^{\frac{3}{4}}}{\sqrt[4]{n\ln n}}\cdot\sqrt[4]{\frac{L_2}{\Theta_1}}\right\}$ | $\min\left\{n^{\frac{3}{2}}L_2\sqrt{\Theta_2},\ \frac{\varepsilon}{\sqrt{n\Theta_2}},\ \frac{\varepsilon^{\frac{3}{4}}}{\sqrt{n}}\cdot\sqrt[4]{\frac{L_2}{\Theta_2}}\right\}$ |
| Calls | $\max\left\{\sqrt{\frac{n\ln nL_2\Theta_1}{\varepsilon}}, \frac{\sigma^2\Theta_1\ln n}{\varepsilon^2}\right\}$ | $\max\left\{\sqrt{\frac{n^2L_2\Theta_2}{\varepsilon}}, \frac{\sigma^2\Theta_2 n}{\varepsilon^2}\right\})$ |

Таблица 1: Параметры Алгоритма 8 для случаев $p=1$ и $p=2$.

Подходящий выбор параметров УРСН приведен в Таблице 1.

Рандомизированный спуск по направлению (РСН) приведен как Алгоритм 9.

**Теорема 2.10.** *Пусть РСН применяется для решения задачи* (42). *Тогда*

$$\mathbb{E}[f(\bar{x}_N)] - f(x_*) \leqslant \frac{384n\rho_nL_2\Theta_p}{N} + \frac{2}{L_2}\frac{\sigma^2}{m} + \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2 + \frac{8\sqrt{2n\Theta_p}}{N}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2, \tag{48}$$

*где* $\Theta_p = V[z_0](x^*)$ *определяется выбранной прокс-функцией и* $\mathbb{E}[\cdot] = \mathbb{E}_{e_1,\ldots,e_N,\xi_{1,1},\ldots,\xi_{N,m}}[\cdot]$.

Подходящий выбор параметров РСН приведен в Таблице 2.

### 2.3.2 Алгоритмы и основные результаты для сильно выпуклых задач

Чтобы получить более быструю скорость сходимости, предположим, что $f$ является $\mu_p$-сильно выпуклой по отношению к $p$-норме. Сделаем также следующее предположение. Пусть $x_*$ зафиксировано и $x$ – случайная точка такая, что $\mathbb{E}_x\left[\|x - x_*\|_p^2\right] \leqslant R_p^2$. Тогда выполнено

$$\mathbb{E}_x d\left(\frac{x - x_*}{R_p}\right) \leqslant \frac{\Omega_p}{2}, \tag{49}$$

где $\mathbb{E}_x$ означает матожидание относительно случайного вектора $x$, а $\Omega_p$ определяется следующим образом. При $p=1$ и нашем выборе прокс-функции (45), $\Omega_p = \mathrm{e}n^{(\kappa-1)(2-\kappa)/\kappa}\ln n = O(\ln n)$ с $\kappa = 1 + \frac{1}{\ln n}$, см. [6, 29].

| | $p=1$ | $p=2$ |
|---|---|---|
| $N$ | $\frac{L_2\Theta_1 \ln n}{\varepsilon}$ | $\frac{nL_2\Theta_2}{\varepsilon}$ |
| $m$ | $\max\left\{1, \frac{\sigma^2}{\varepsilon L_2}\right\}$ | $\max\left\{1, \frac{\sigma^2}{\varepsilon L_2}\right\}$ |
| $\Delta_\zeta$ | $\min\left\{\frac{(\ln n)^2}{n} L_2^2\Theta_1,\ \frac{\varepsilon^2}{n\Theta_1},\ \frac{\varepsilon L_2}{n}\right\}$ | $\min\left\{nL_2^2\Theta_2,\ \frac{\varepsilon^2}{n\Theta_2},\ \frac{\varepsilon L_2}{n}\right\}$ |
| $\Delta_\eta$ | $\min\left\{\frac{\ln n}{\sqrt{n}} L_2\sqrt{\Theta_1},\ \frac{\varepsilon}{\sqrt{n\Theta_1}},\ \sqrt{\frac{\varepsilon L_2}{n}}\right\}$ | $\min\left\{\sqrt{n}L_2\sqrt{\Theta_2},\ \frac{\varepsilon}{\sqrt{n\Theta_2}},\ \sqrt{\frac{\varepsilon L_2}{n}}\right\}$ |
| $Nm$ | $\max\left\{\frac{L_2\Theta_1 \ln n}{\varepsilon}, \frac{\sigma^2\Theta_1 \ln n}{\varepsilon^2}\right\}$ | $\max\left\{\frac{nL_2\Theta_2}{\varepsilon}, \frac{n\sigma^2\Theta_2}{\varepsilon^2}\right\}$ |

Таблица 2: Параметры Алгоритма 9 для случаев $p=1$ и $p=2$.

Для $p=2$ и соответствующего выбора прокс-функции, $\Omega_p = 1$. Ускоренный рандомизированный спуск по направлению для сильно выпуклых задач (УРСНсв) приведен как Алгоритм 10.

**Algorithm 10** Ускоренный рандомизированный спуск по направлению для сильно выпуклых задач (УРСНсв)

**Вход:** $x_0$ —начальная точка, удовлетворяющая $\|x_0 - x_*\|_p^2 \leq R_p^2$; $K \geqslant 1$ — число итераций; $\mu_p$ – параметр сильной выпуклости.

**Выход:** точка $u_K$.

1: Положить $N_0 = \left\lceil \sqrt{\frac{8aL_2\Omega_p}{\mu_p}} \right\rceil$, где $a = 384n^2\rho_n$.
2: **for** $k = 0, \ldots, K-1$ **do**
3: $\quad m_k := \max\left\{1, \left\lceil \frac{32\sigma^2 N_0 2^k}{nL_2\mu_p R_p^2} \right\rceil \right\}, \quad R_k^2 := R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right),$
4: $\quad$ Положить $d_k(x) = R_k^2 d\left(\frac{x-u_k}{R_k}\right)$.
5: $\quad$ Запустить УРСН из точки $u_k$ с прокс-функцией $d_k(x)$ на $N_0$ шагов с размером батча $m_k$.
6: $\quad$ Положить $u_{k+1} = y_{N_0}$, $k = k+1$.
7: **end for**
8: **return** $u_K$

**Теорема 2.11.** *Пусть $f$ в задаче (42) является $\mu_p$-сильно выпуклой и УРСНсв применяется для ее решения. Тогда*

$$\mathbb{E}f(u_K) - f^* \leqslant \frac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta. \tag{50}$$

*где* $\Delta = \frac{61N_0}{24L_2}\Delta_\zeta + \frac{122N_0}{3L_2}\Delta_\eta^2 + \frac{12\sqrt{2nR_p^2\Omega_p}}{N_0^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2$.
*Кроме того, если $\Delta_\zeta$ и $\Delta_\eta$ выбраны так, что $2\Delta \leqslant \varepsilon/2$, то оракульная*

*сложность для получения $\varepsilon$-решения равна*

$$\widetilde{O}\left(\max\left\{n^{\frac{1}{2}+\frac{1}{q}}\sqrt{\frac{L_2\Omega_p}{\mu_p}}\log_2\frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right).$$

Подходящие значения параметров в алгоритме УРСНсв приведены в Таблице 3.

| | $p=1$ | $p=2$ |
|---|---|---|
| $\Delta_\zeta$ | $\min\left\{\varepsilon\sqrt{\frac{L_2\mu_1}{n\ln n\Omega_1}}, \varepsilon^2\frac{n(\ln n)^2L_2^2\Omega_1}{R_1^2\mu_1^2}, \varepsilon\cdot\frac{\mu_1}{n\Omega_1}\right\}$ | $\min\left\{\varepsilon\sqrt{\frac{L_2\mu_2}{n^2\Omega_2}}, \varepsilon^2\frac{n^3L_2^2\Omega_2}{R_2^2\mu_2^2}, \varepsilon\cdot\frac{\mu_2}{n\Omega_2}\right\}$ |
| $\Delta_\eta$ | $\min\left\{\sqrt{\varepsilon}\sqrt[4]{\frac{L_2\mu_1}{n\ln n\Omega_1}}, \varepsilon\frac{\sqrt{n}\ln nL_2\sqrt{\Omega_1}}{R_1\mu_1}, \sqrt{\varepsilon}\cdot\sqrt{\frac{\mu_1}{n\Omega_1}}\right\}$ | $\min\left\{\sqrt{\varepsilon}\sqrt[4]{\frac{L_2\mu_2}{n^2\Omega_2}}, \varepsilon\frac{\sqrt{n^3}L_2\sqrt{\Omega_2}}{R_2\mu_2}, \sqrt{\varepsilon}\cdot\sqrt{\frac{\mu_2}{n\Omega_2}}\right\}$ |
| Calls | $\max\left\{\sqrt{\frac{n\ln nL_2\Omega_1}{\mu_1}}\log_2\frac{\mu_1R_1^2}{\varepsilon}, \frac{\sigma^2\Omega_1\ln n}{\mu_1\varepsilon}\right\}$ | $\max\left\{n\sqrt{\frac{L_2\Omega_2}{\mu_2}}\log_2\frac{\mu_2R_2^2}{\varepsilon}, \frac{n\sigma^2\Omega_2}{\mu_2\varepsilon}\right\}$ |

Таблица 3: Параметры Алгоритма 10 для случаев $p=1$ и $p=2$.

Рандомизированный спуск по направлению для сильно выпуклых задач (РСНсв) приведен как Алгоритм 11.

**Algorithm 11** Рандомизированный спуск по направлению для сильно выпуклых задач (РСНсв)

**Вход:** $x_0$ —начальная точка, удовлетворяющая $\|x_0-x_*\|_p^2\leq R_p^2$; $K\geqslant 1$ — число итераций; $\mu_p$ – параметр сильной выпуклости.

**Выход:** Точка $u_K$.

1: Положить $N_0=\left\lceil\frac{8aL_2\Omega_p}{\mu_p}\right\rceil$, где $a=384n\rho_n$.
2: **for** $k=0,\ldots,K-1$ **do**
3: $m_k:=\max\left\{1,\left\lceil\frac{16\sigma^2 2^k}{L_2\mu_pR_p^2}\right\rceil\right\}$, $R_k^2:=R_p^2 2^{-k}+\frac{4\Delta}{\mu_p}\left(1-2^{-k}\right)$,
4: Положить $d_k(x)=R_k^2d\left(\frac{x-u_k}{R_k}\right)$.
5: Запустить РСН из точки $u_k$ с прокс-функцией $d_k(x)$ на $N_0$ шагов с размером батча $m_k$.
6: Положить $u_{k+1}=y_{N_0}$, $k=k+1$.
7: **end for**
8: **return** $u_K$

**Теорема 2.12.** *Пусть $f$ в задаче (42) является $\mu_p$-сильно выпуклой и РСНсв применяется для ее решения. Тогда*

$$\mathbb{E}f(u_K)-f^*\leqslant\frac{\mu_pR_p^2}{2}\cdot 2^{-K}+2\Delta. \tag{51}$$

*где* $\Delta = \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2 + \frac{8\sqrt{2nR_p^2\Omega_p}}{N_0}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2$. *Кроме того, если* $\Delta_\zeta$ *и* $\Delta_\eta$ *выбраны так, что* $2\Delta \leqslant \varepsilon/2$, *то оракульная сложность для получения* $\varepsilon$*-решения равна*

$$\widetilde{O}\left(\max\left\{\frac{n^{\frac{2}{q}}L_2\Omega_p}{\mu_p}\log_2\frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right).$$

Подходящие значения параметров для РСНсв приведены в Таблице 4.

| | $p=1$ | $p=2$ |
|---|---|---|
| $\Delta_\zeta$ | $\min\left\{\frac{\varepsilon L_2}{n}, \varepsilon^2\frac{(\ln n)^2L_2^2}{nR_1^2\mu_1^2}, \varepsilon\frac{\mu_1}{n\Omega_1}\right\}$ | $\min\left\{\frac{\varepsilon L_2}{n}, \varepsilon^2\frac{nL_2^2}{R_2^2\mu_2^2}, \varepsilon\frac{\mu_2}{n\Omega_2}\right\}$ |
| $\Delta_\eta$ | $\min\left\{\sqrt{\frac{\varepsilon L_2}{n}}, \varepsilon\frac{\ln nL_2}{\sqrt{n}R_1\mu_1}, \sqrt{\varepsilon\frac{\mu_1}{n\Omega_1}}\right\}$ | $\min\left\{\sqrt{\frac{\varepsilon L_2}{n}}, \varepsilon\frac{\sqrt{n}L_2}{R_2\mu_2}, \sqrt{\varepsilon\frac{\mu_2}{n\Omega_2}}\right\}$ |
| Calls | $\max\left\{\frac{L_2\Omega_1\ln n}{\mu_1}\log_2\frac{\mu_1R_1^2}{\varepsilon}, \frac{\sigma^2\Omega_1}{\mu_1\varepsilon}\right\}$ | $\max\left\{\frac{nL_2\Omega_2}{\mu_2}\log_2\frac{\mu_2R_2^2}{\varepsilon}, \frac{n\sigma^2\Omega_2}{\mu_2\varepsilon}\right\}$ |

Таблица 4: Параметры Алгоритма 11 для случаев $p=1$ и $p=2$.

# 3 Прямо-двойственные методы

В этом разделе рассматриваются предложенные прямо-двойственные методы первого порядка для выпуклых задач с линейными ограничениями.

## 3.1 Прямо-двойственные методы для решения бесконечномерных игр

Результаты этого раздела опубликованы в статье [30]. Рассмотрим два движущихся объекта с динамикой, заданной уравнениями

$$\begin{aligned}&\dot{x}(t) = A_x(t)x(t) + B(t)u(t), \dot{y}(t) = A_y(t)y(t) + C(t)v(t),\\&(x(0), y(0)) = (x_0, y_0).\end{aligned} \tag{52}$$

Здесь $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^m$ – фазовые векторы этих объектов, $u(t)$ – управление первого объекта (преследователь), $v(t)$ – управление второго объекта (убегающий). Матрицы $A_x(t), A_y(t), B(t), C(t)$ – непрерывные функции времени, имеющие соответствующие размеры. Система рассматривается на интервале $[0, \theta]$. Управления подчинены ограничениям следующего вида $u(t) \in P \subseteq \mathbb{R}^p$, $v(t) \in Q \subseteq \mathbb{R}^q \quad \forall t \in [0, \theta]$. Предполагается, что $P, Q$ – замкнутые выпуклые множества.

Цель преследователя состоит в минимизации функционала

$$F(u,v)+\Phi(x(\theta),y(\theta)):=\int_0^\theta \tilde{F}(\tau,u(\tau),v(\tau))d\tau+\Phi(x(\theta),y(\theta)). \qquad (53)$$

Цель убегающего противоположная. Необходимо найти оптимальный гарантированный результат для каждого объекта, что приводит к задаче поиска седловой точки этого функционала. Делаются следующие предположения

- $u(\cdot)\in L^2([0,\theta],\mathbb{R}^p)$, и $v(\cdot)\in L^2([0,\theta],\mathbb{R}^q)$ (для простоты обозначим $L^2([0,\theta],\mathbb{R}^p)$ как $L^2_p$ и $L^2([0,\theta],\mathbb{R}^q)$ как $L^2_q$),
- существует седловая точка в рассматриваемом классе стратегий,
- функция $F(u,v)$ полунепрерывна сверху по $v$ и полунепрерывна снизу по $u$,
- $\Phi(x,y)$ непрерывна.

Обозначим через $V_x(t,\tau)$ матрицу перехода для состояния в первом уравнении (52). Эта матрица является единственным решением матричной задачи Коши

$$\frac{dV_x(t,\tau)}{dt}=A_x(t)V_x(t,\tau),\quad t\geq\tau,\quad V_x(\tau,\tau)=E.$$

Здесь $E$ обозначает единичную матрицу. Если матрица $A_x(t)$ постоянна, то $V_x(t,\tau)=e^{(t-\tau)A}$.

Если разрешить первое уравнение в (52), то можно выразить $x(\theta)$ как результат применения линейного оператора $\mathcal{B}:L^2_p\to\mathbb{R}^n$:

$$x(\theta)=V_x(\theta,0)x_0+\int_0^\theta V_x(\theta,\tau)B(\tau)u(\tau)d\tau:=\tilde{x}_0+\mathcal{B}u. \qquad (54)$$

Ниже также используется сопряженный к нему оператор $\mathcal{B}^*$, который можно выразить явно. Пусть $\mu$ – $n$-мерный вектор. Тогда

$$\langle\mu,\mathcal{B}u\rangle=\langle\mu,\int_0^\theta V_x(\theta,\tau)B(\tau)u(\tau)d\tau\rangle=\int_0^\theta\langle\mu,V_x(\theta,\tau)B(\tau)u(\tau)\rangle d\tau=$$
$$=\int_0^\theta\langle B^T(\tau)V_x^T(\theta,\tau)\mu,u(\tau)\rangle d\tau=\langle\mathcal{B}^*\mu,u\rangle.$$

Отметим, что вектор $\zeta(t)=V_x^T(\theta,t)\mu$ является решением следующей задачи Коши:

$$\dot{\zeta}(t)=-A_x^T(t)\zeta(t),\quad \zeta(\theta)=\mu,\quad t\in[0,\theta].$$

Таким образом, можно решить это дифференциальное уравнение и найти $\mathcal{B}^*\mu$ как $\mathcal{B}^*\mu(t) = B^T(t)\zeta(t)$.

Аналогично вводится матрица перехода $V_y(t,\tau)$ для второго уравнения в (52), задается оператор $\mathcal{C} : L_q^2 \to \mathbb{R}^m$ с помощью равенства $\mathcal{C}v := \int_0^\theta V_y(\theta,\tau)C(\tau)v(\tau)d\tau$ и вектор $\tilde{y}_0 := V_y(\theta,0)y_0$. Сопряженный оператор $\mathcal{C}^*$ также может быть вычислен с помощью решения дифференциального уравнения.

Итак, рассматривается дифференциальная игра следующего вида

$$\min_{u\in\mathcal{U}}\left[\max_{v\in\mathcal{V}}\{F(u,v)+\Phi(x,y) : y=\tilde{y}_0+\mathcal{C}v\} : x=\tilde{x}_0+\mathcal{B}u\right], \tag{55}$$

где

$$\mathcal{U} := \{u(\cdot)\in L_p^2 : u(t)\in P \quad \forall t\in[0,\theta]\}, \mathcal{V} := \{v(\cdot)\in L_q^2 : v(t)\in Q \quad \forall t\in[0,\theta]\}$$

– множества допустимых стратегий игроков и $u\in\mathcal{U}$, $v\in\mathcal{V}$ означает, что $u(\cdot)\in\mathcal{U}$, $v(\cdot)\in\mathcal{V}$. Ниже предлагается численный метод для вычисления приближенного решения этой задачи (55) при двух предположениях

**A1** Множества $P$ и $Q$ ограничены.

**A2** В (53) функционал $F(\cdot,v)$ выпуклый при любом фиксированном $v$, функционал $F(u,\cdot)$ вогнутый при любом фиксированном $u$, $\Phi(\cdot,y)$ выпукла при любом фиксированном $y$, $\Phi(x,\cdot)$ вогнута при любом фиксированном $x$.

Из **A1**, так как операторы $\mathcal{B},\mathcal{C}$ ограничены, следует, что $x(\theta), y(\theta)$ тоже ограничены и можно эквивалентно преобразовать задачу (55) следующим образом

$$\min_{u\in\mathcal{U},x\in X}\left[\max_{v\in\mathcal{V},y\in Y}\{F(u,v)+\Phi(x,y) : y=\tilde{y}_0+\mathcal{C}v\} : x=\tilde{x}_0+\mathcal{B}u\right] =$$
$$\max_{v\in\mathcal{V},y\in Y}\left[\min_{u\in\mathcal{U},x\in X}\{F(u,v)+\Phi(x,y) : x=\tilde{x}_0+\mathcal{B}u\} : y=\tilde{y}_0+\mathcal{C}v\right], \tag{56}$$

где множества $X$ и $Y$ замкнуты, выпуклы и ограничены. Введем пространства двойственных переменных $\lambda\in\mathbb{R}^m$ и $\mu\in\mathbb{R}^n$ соответствующих линейным ограничениям в задаче (56) и нормы $\|\cdot\|_\lambda$ и $\|\cdot\|_\mu$ на этих пространствах. Определим нормы в сопряженном пространстве стандартным образом

$$\|s_\lambda\|_{\lambda,*} := \max\{\langle s_\lambda,\lambda\rangle : \|\lambda\|_\lambda\le 1\}, \quad \|s_\mu\|_{\mu,*} := \max\{\langle s_\mu,\mu\rangle : \|\mu\|_\mu\le 1\}.$$

**Лемма 3.1.** *Пусть выполнены предположения* **A1**, **A2**. *Также предположим, что $F(u,v)$ полунепрерывна сверху по $v$ и полунепрерывна снизу по $u$, функция $\Phi(x,y)$ непрерывна, а также, что множества $P$ и $Q$ выпуклы и замкнуты. Тогда задача* (56) *эквивалентна задаче*

$$\begin{aligned}&\min_\lambda\max_\mu\{\min_{u\in\mathcal{U}}\max_{v\in\mathcal{V}}[F(u,v)-\langle\mu,\mathcal{B}u\rangle+\langle\lambda,\mathcal{C}v\rangle]\\ &+\min_{x\in X}\max_{y\in Y}[\Phi(x,y)+\langle\mu,x\rangle-\langle\lambda,y\rangle]-\langle\mu,\tilde{x}_0\rangle+\langle\lambda,\tilde{y}_0\rangle\},\end{aligned} \tag{57}$$

*которую будем назвать сопряженной задачей к задаче* (56).

Обозначим $\psi(\lambda, \mu)$ функцию, для которой в (57) ищется седловая точка.

### 3.1.1 Алгоритм для выпукло-вогнутой задачи

Пусть задана прокс-фцнкция $d_\lambda(\lambda)$ с прокс-цетром $\lambda_0$, которая является сильно выпуклой с параметром $\sigma_\lambda$ в заданной норме $\|\cdot\|_\lambda$. Для $\mu$ делаются аналогичные предположения. Так как $(\lambda^*, \mu^*)$ – седловая точка, то $(\lambda^*, \mu^*)$ является слабым решением следующего вариационного неравенства $\langle g(\lambda, \mu), (\lambda - \lambda^*, \mu - \mu^*)\rangle \geq 0, \quad \forall \lambda, \mu$, где $g(\lambda, \mu) := (\psi'_\lambda(\lambda, \mu), -\psi'_\mu(\lambda, \mu))$. Применим метод простых двойственных усреднений (ПДУ) из [31] для поиска приближенного решения конечномерной сопряженной задачи (57). Выберем некоторое $\kappa \in ]0, 1[$. Рассмотрим пространство $z := (\lambda, \mu)$ с нормой

$$\|z\|_z := \sqrt{\kappa\sigma_\lambda \|\lambda\|_\lambda^2 + (1-\kappa)\sigma_\mu \|\mu\|_\mu^2}, \tag{58}$$

оракул $g(z) := (g_\lambda(z), -g_\mu(z))$, новую прокс-функцию
$d(z) := \kappa d_\lambda(\lambda) + (1 - \kappa)\, d_\mu(\mu)$, являющуюся сильно выпуклой с параметром $\sigma_0 = 1$ по отношению к норме (58). Определим $W := \mathbb{R}^m \times \mathbb{R}^n$. Сопряженная к (58) норма определяется как

$$\|g\|_{z,*} := \sqrt{\frac{1}{\kappa\sigma_\lambda}\|g_\lambda\|_{\lambda,*}^2 + \frac{1}{(1-\kappa)\sigma_\mu}\|g_\mu\|_{\mu,*}^2}.$$

Отметим, что в силу ограниченности доступна следующая равномерная оценка сверху для ответов оракула: $\|g(\lambda, \mu)\|_{z,*}^2 \leq L^2 := \frac{L_\lambda^2}{\kappa\sigma_\lambda} + \frac{L_\mu^2}{(1-\kappa)\sigma_\mu}$, где $L_\lambda := \sqrt{\theta}\,\|\mathcal{C}\|_{\lambda, L_q^2}\,\mathrm{diam}_2 Q + \mathrm{diam}_{\lambda,*} Y + \|\tilde{y}_0\|_{\lambda,*}$ и $L_\mu := \sqrt{\theta}\,\|\mathcal{B}\|_{\mu, L_p^2}\,\mathrm{diam}_2 P + \mathrm{diam}_{\mu,*} X + \|\tilde{x}_0\|_{\mu,*}$.

Алгоритм ПДУ для задачи (57) выглядит следующим образом

1. Инициализация: Положить $s_0 = 0$. Выбрать $z_0$, $\gamma > 0$.

2. Итерация ($k \geq 0$):

$$\begin{aligned} &\text{Вычислить } g_k = g(z_k). \text{ Положить } s_{k+1} = s_k + g_k. \\ &\beta_{k+1} = \gamma\hat{\beta}_{k+1}. \text{ Положить } z_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1}). \end{aligned} \tag{M1}$$

Здесь последовательность $\hat{\beta}_{k+1}$ определена рекуррентно как $\hat{\beta}_0 = \hat{\beta}_1 = 1$, $\hat{\beta}_{i+1} = \hat{\beta}_i + \frac{1}{\hat{\beta}_i}$, при $i \geq 1$. Отображение $\pi_\beta(s)$ определяется как $\pi_\beta(s) := \arg\min_{z \in W} \{-\langle s, z\rangle + \beta d(z)\}$.

Выберем $D_\lambda, D_\mu$ так, что $d_\lambda(\lambda_i) \leq D_\lambda$, $d_\mu(\mu_i) \leq D_\mu$ для всех $i \geq 0$ и пара $(\lambda^*, \mu^*)$ является внутренним решением:

$$\mathfrak{B}^\lambda_{r/\sqrt{\kappa\sigma_\lambda}}(\lambda^*) \subseteq W_\lambda := \{\, \lambda : \, d_\lambda(\lambda) \leq D_\lambda\},$$

$$\mathfrak{B}^{\mu}_{r/\sqrt{(1-\kappa)\sigma_\mu}}(\mu^*) \subseteq W_\mu := \{ \mu : d_\mu(\mu) \leq D_\mu \}$$

для некоторого $r > 0$. Тогда $z^* := (\lambda^*, \mu^*) \in \mathcal{F}_D := \{ z \in W : d(z) \leq D \}$ с $D := \kappa D_\lambda + (1-\kappa) D_\mu$ и $\mathfrak{B}^z_r(z^*) \subseteq \mathcal{F}_D$.

Введем функцию зазора

$$\delta_k(D) := \max_z \left\{ \sum_{i=0}^{k} \langle g_i, z_i - z \rangle : z \in \mathcal{F}_D \right\}. \tag{59}$$

Из Теоремы 2 [31] следует, что

$$\frac{1}{k+1}\delta_k(D) \leq \frac{\hat{\beta}_{k+1}}{k+1}\left(\gamma D + \frac{L^2}{2\gamma}\right). \tag{60}$$

Обозначим

$$(\hat{u}_{k+1}, \hat{v}_{k+1}, \hat{x}_{k+1}, \hat{y}_{k+1}) := \frac{1}{k+1}\sum_{i=0}^{k}(u_i, v_i, x_i, y_i), \tag{61}$$

где $(u_i, v_i)$, $(x_i, y_i)$ – седловые точки, соответствующие $(\lambda_i, \mu_i)$ в (57). Определим функцию

$$\begin{aligned}&\phi(u,x,v,y) := \min_\lambda \max_\mu \{F(u,v) + \Phi(x,y) + \langle \mu, x - \tilde{x}_0 - \mathcal{B}u\rangle + \\ &+\langle \lambda, \mathcal{C}v + \tilde{y}_0 - y\rangle : d_\lambda(\lambda) \leq D_\lambda, d_\mu(\mu) \leq D_\mu\}.\end{aligned} \tag{62}$$

Так как $d_\lambda(\lambda^*) \leq D_\lambda$, $d_\mu(\mu^*) \leq D_\mu$ и сопряженная задача эквивалентна исходной, то исходная задача эквивалентна следующей задаче

$$\min_{u\in\mathcal{U}, x\in X} \max_{v\in\mathcal{V}, y\in Y} \phi(u,x,v,y). \tag{63}$$

Введем две вспомогательные функции

$$\xi(u,x) := \max_{v\in\mathcal{V}, y\in Y} \phi(u,x,v,y), \tag{64}$$

$$\eta(v,y) := \min_{u\in\mathcal{U}, x\in X} \phi(u,x,v,y). \tag{65}$$

Заметим, что $\xi(u,x)$ выпукла, $\eta(v,y)$ вогнута и $\xi(u,x) \geq \phi(u^*,x^*,v^*,y^*) \geq \eta(v,y)$ для всех $u \in \mathcal{U}, v \in \mathcal{V}, x \in X, y \in Y$, где $\phi(u^*,x^*,v^*,y^*)$ – решение (63).

**Теорема 3.1.** *Пусть выполнены предположения* **A1** *и* **A2**. *Тогда точки* (61)*, генерируемые методом* (M1)*, удовлетворяют соотношениям*

$$\xi(\hat{u}_{k+1}, \hat{x}_{k+1}) - \eta(\hat{v}_{k+1}, \hat{y}_{k+1}) \leq \frac{\hat{\beta}_{k+1}}{k+1}\left(\gamma D + \frac{L^2}{2\gamma}\right), \tag{66}$$

$$\begin{aligned}\|\tilde{x}_0 + \mathcal{B}\hat{u}_{k+1} - \hat{x}_{k+1}\|_{\mu,*} &\leq \tfrac{\hat{\beta}_{k+1}\sqrt{\sigma_\mu}}{r(k+1)}\left(\gamma D + \tfrac{L^2}{2\gamma}\right), \\ \|\tilde{y}_0 + \mathcal{C}\hat{v}_{k+1} - \hat{y}_{k+1}\|_{\lambda,*} &\leq \tfrac{\hat{\beta}_{k+1}\sqrt{\sigma_\lambda}}{r(k+1)}\left(\gamma D + \tfrac{L^2}{2\gamma}\right).\end{aligned} \tag{67}$$

### 3.1.2 Алгоритм для сильно выпукло-вогнутых задач

В этом разделе рассматривается задача (55) при более сильных предположениях, что позволяет получить более высокую скорость сходимости.

Сделаем следующие предположения.

**A3** Функция $F(\cdot, v)$ сильно выпукла при фиксированном $v$ с константой $\sigma_{F_u}$, которая не зависит от $v$, функция $F(u, \cdot)$ сильно вогнута при фиксированном $u$ с константой $\sigma_{F_v}$, которая не зависит от $u$. Также справедливы неравенства

$$\|\nabla_u F(u, v_1) - \nabla_u F(u, v_2)\|_{L_p^2} \leq L_{uv} \|v_1 - v_2\|_{L_q^2}, \tag{68}$$

$$\|\nabla_v F(u_1, v) - \nabla_v F(u_2, v)\|_{L_q^2} \leq L_{vu} \|u_1 - u_2\|_{L_p^2}. \tag{69}$$

**A4** $\Phi(\cdot, y)$ сильно выпукла при фиксированном $y$ с константой $\sigma_{\Phi x}$, которая не зависит от $y$, $\Phi(x, \cdot)$ сильно вогнута при фиксированном $x$ с константой $\sigma_{\Phi y}$, которая не зависит от $x$. Также справедливы неравенства

$$\|\nabla_x \Phi(x, y_1) - \nabla_x \Phi(x, y_2)\|_\mu \leq L_{xy} \|y_1 - y_2\|_{\lambda,*}, \tag{70}$$

$$\|\nabla_y \Phi(x_1, y) - \nabla_y \Phi(x_2, y)\|_\lambda \leq L_{yx} \|x_1 - x_2\|_{\mu,*}, \tag{71}$$

$$\|\nabla_x \Phi(x_1, y) - \nabla_x \Phi(x_2, y)\|_\mu \leq L_{xx} \|x_1 - x_2\|_{\mu,*}, \tag{72}$$

$$\|\nabla_y \Phi(x, y_1) - \nabla_y \Phi(x, y_2)\|_\lambda \leq L_{yy} \|y_1 - y_2\|_{\lambda,*}. \tag{73}$$

Аналогично Лемме 3.1 получаем следующую сопряженную задачу для (55)

$$\begin{aligned} \min_\lambda \max_\mu \{ \quad & \min_{u\in\mathcal{U}} \max_{v\in\mathcal{V}} [F(u, v) - \langle \mu, \mathcal{B}u\rangle + \langle \lambda, \mathcal{C}v\rangle] \\ & + \min_x \max_y [\Phi(x, y) + \langle \mu, x\rangle - \langle \lambda, y\rangle] - \langle \mu, \tilde{x}_0\rangle + \langle \lambda, \tilde{y}_0\rangle \}. \end{aligned} \tag{74}$$

Предположим, что нормы $\|\cdot\|_\lambda$ и $\|\cdot\|_\mu$ Евклидовы. Введем прокс-функцию $d_\lambda(\lambda) := \frac{\sigma_\lambda}{2} \|\lambda\|_\lambda^2$. Функция $d_\lambda(\lambda)$ сильно выпукла с параметром $\sigma_\lambda$. Для переменной $\mu$ введем прокс-функцию $d_\mu(\mu) := \frac{\sigma_\mu}{2} \|\mu\|_\mu^2$, которая сильно выпукла с параметром $\sigma_\mu$ в норме $\|\cdot\|_\mu$.

Для любых $\lambda_1, \lambda_2 \in \mathbb{R}^m$ определим дивергенцию Брегмана

$$\omega_\lambda(\lambda_1, \lambda_2) := d_\lambda(\lambda_2) - d_\lambda(\lambda_1) - \langle \nabla d_\lambda(\lambda_1), \lambda_2 - \lambda_1\rangle.$$

Используя явное выражение для $d_\lambda(\lambda)$, получим $\omega_\lambda(\lambda_1, \lambda_2) = \frac{\sigma_\lambda}{2} \|\lambda_1 - \lambda_2\|^2$. Выберем $\bar{\lambda} = 0$ в качестве центра. Тогда $\omega_\lambda(\bar{\lambda}, \lambda) = d_\lambda(\lambda)$. Для $\mu$ введем аналогичные объекты.

Поиск седловой точки $(\lambda^*, \mu^*)$ для сопряженной задачи (74) эквивалентен решению вариационного неравенства

$$\langle g(\lambda, \mu), (\lambda - \lambda^*, \mu - \mu^*)\rangle \geq 0, \quad \forall \lambda, \mu, \tag{75}$$

$$\text{где} \quad g(\lambda,\mu) := (\nabla_\lambda \psi(\lambda,\mu), -\nabla_\mu \psi(\lambda,\mu)). \tag{76}$$

Выберем некоторое $\kappa \in ]0,1[$. Рассмотрим пространство векторов $z := (\lambda,\mu)$ с нормой

$$\|z\|_z := \sqrt{\kappa\sigma_\lambda \|\lambda\|_\lambda^2 + (1-\kappa)\sigma_\mu \|\mu\|_\mu^2},$$

оракул $g(z) := (\nabla_\lambda \psi(\lambda,\mu), -\nabla_\mu \psi(\lambda,\mu))$, новую прокс-функцию

$$d(z) := \kappa d_\lambda(\lambda) + (1-\kappa) d_\mu(\mu),$$

которая сильно выпукла с константой $\sigma_0 = 1$. Определим $W := \mathbb{R}^m \times \mathbb{R}^n$, дивергенцию Брегмана

$$\omega(z_1, z_2) := \kappa\omega_\lambda(\lambda_1,\lambda_2) + (1-\kappa)\omega_\lambda(\mu_2,\mu_2),$$

которая имеет явный вид $\omega(z_1,z_2) = d(z_1 - z_2)$ и прокс-центр $\bar{z} = (0,0)$. Тогда, $\omega(\bar{z},z) = d(z)$. Норма в двойственном пространстве определяется как

$$\|g\|_{z,*} := \sqrt{\frac{1}{\kappa\sigma_\lambda}\|g_\lambda\|_{\lambda,*}^2 + \frac{1}{(1-\kappa)\sigma_\mu}\|g_\mu\|_{\mu,*}^2}.$$

Согласно [32] для решения (75) можно использовать следующий метод

1. Инициализация: Зафиксировать $\beta = L$ (константа Липшица $g$) . Положить $s_{-1} = 0$.

2. Итерация ($k \geq 0$):

$$\begin{aligned} &\text{Вычислить } x_k = T_\beta(\bar{z}, s_{k-1}), \\ &\text{Вычислить } z_k = T_\beta(x_k, -g(x_k)), \\ &\text{Положить } s_k = s_{k-1} - g(z_k). \end{aligned} \tag{M2}$$

Здесь $T_\beta(z,s) := \arg\max_{x\in W}\{\langle s, x - z\rangle - \beta\omega(z,x)\}$.

Аналогично [31] можно показать, что метод (M2) генерирует ограниченную последовательность $\{z_i\}_{i\geq 0}$. Следовательно, последовательности $\{\lambda_i\}_{i\geq 0}, \{\mu_i\}_{i\geq 0}$ также ограничены. Так как седловая точка в задаче (55) существует, то существует седловая точка $(\lambda^*,\mu^*)$ для сопряженной задачи (74). Поэтому можно выбрать $D_\lambda, D_\mu$ такое, что $d_\lambda(\lambda_i) \leq D_\lambda$, $d_\mu(\mu_i) \leq D_\mu$ для всех $i \geq 0$, что также гарантирует, что $(\lambda^*,\mu^*)$ является внутренним решением:

$$\mathfrak{B}^\lambda_{r/\sqrt{\kappa\sigma_\lambda}}(\lambda^*) \subseteq W_\lambda := \{\lambda : d_\lambda(\lambda) \leq D_\lambda\},$$

$$\mathfrak{B}^\mu_{r/\sqrt{(1-\kappa)\sigma_\mu}}(\mu^*) \subseteq W_\mu := \{\mu : d_\mu(\mu) \leq D_\mu\}$$

для некоторого $r > 0$. Тогда $z^* := (\lambda^*,\mu^*) \in \mathcal{F}_D := \{z \in W : d(z) \leq D\}$ с $D := \kappa D_\lambda + (1-\kappa)D_\mu$ и $\mathfrak{B}^z_r(z^*) \subseteq \mathcal{F}_D$.

**Теорема 3.2.** *Пусть выполнены предположения* **A3** *и* **A4**, $\kappa = \frac{\sigma_\mu}{\sigma_\mu+\sigma_\lambda}$, *и*

$$L = \frac{\sigma_\lambda+\sigma_\mu}{\sigma_\mu\sigma_\lambda}\sqrt{2\left(\frac{\|\mathcal{C}\|^2_{\lambda,L_q^2}}{\sigma_{F_v}} + \frac{1}{\sigma_{\Phi_y}} + \frac{\|\mathcal{B}\|_{\mu,L_p^2}\|\mathcal{C}\|_{\lambda,L_q^2}L_{vu}}{\sigma_{F_u}\sigma_{F_v}} + \frac{L_{yx}}{\sigma_{\Phi x}\sigma_{\Phi_y}}\right)} \Big/ \sqrt{\left(\frac{\|\mathcal{B}\|_{\mu,L_p^2}\|\mathcal{C}\|_{\lambda,L_q^2}L_{uv}}{\sigma_{F_u}\sigma_{F_v}} + \frac{L_{xy}}{\sigma_{\Phi_x}\sigma_{\Phi y}} + \frac{\|\mathcal{B}\|^2_{\mu,L_p^2}}{\sigma_{F_u}} + \frac{1}{\sigma_{\Phi_x}}\right)}. \tag{77}$$

*Пусть точки* $z_i = (\lambda_i, \mu_i), i \geq 0$ *сгенерированы методом* (M2). *Пусть точки в* (61) *определены точками* $(u_i, v_i)$, $(x_i, y_i)$, *которые являются седловыми точками, соответствующими* $(\lambda_i, \mu_i)$ *в* (74). *Тогда для функций* $\xi(u, x), \eta(v, y)$, *определенных в* (64) *и* (65), *справедливо*

$$\xi(\hat{u}_{k+1}, \hat{x}_{k+1}) - \eta(\hat{v}_{k+1}, \hat{y}_{k+1}) \leq \frac{LD}{k+1}. \tag{78}$$

*Также выполнено*

$$\|\mathcal{B}\hat{u}_{k+1} + \tilde{x}_0 - \hat{x}_{k+1}\|_{\mu,*} \leq \frac{LD\sqrt{\sigma_\mu}}{r(k+1)}, \quad \|\mathcal{C}\hat{v}_{k+1} + \tilde{y}_0 - \hat{y}_{k+1}\|_{\lambda,*} \leq \frac{LD\sqrt{\sigma_\lambda}}{r(k+1)}.$$

## 3.2 Ускоренный прямо-двойственный градиентный метод для сильно выпуклых задач с линейными ограничениями

Результаты этого раздела опубликованы в статьях [33, 34].

Основной мотивацией для алгоритмов этого раздела является приближенное вычисление оптимально-транспортного расстояния (OT-расстояния), которое состоит в приближенном решении задачи оптимального транспорта (OT) [35]:

$$\min_{X \in \mathcal{U}(r,c)} \langle C, X\rangle,$$
$$\mathcal{U}(r, c) := \{X \in \mathbb{R}_+^{n\times n} : X\mathbf{1} = r,\ X^T\mathbf{1} = c\}, \tag{79}$$

где $X$ – *транспортный план*, $C \in \mathbb{R}_+^{n\times n}$ – заданная матрица затрат, $r, c \in \mathbb{R}^n$ – заданные векторы из вероятностного симплекса $\Delta^n$, $\mathbf{1}$ – вектор со всеми компонентами равными 1. Регуляризованная задача OT ставится следующим образом

$$\min_{X \in \mathcal{U}(r,c)} \langle C, X\rangle + \gamma\mathcal{R}(X), \tag{80}$$

где $\gamma > 0$ – параметр регуляризации и $\mathcal{R}(X)$ – сильно выпуклый регуляризатор, например, минус энтропия или квадрат Евклидовой нормы. Цель состоит в поиске такого $\widehat{X} \in \mathcal{U}(r, c)$, что

$$\langle C, \widehat{X}\rangle \leq \min_{X \in \mathcal{U}(r,c)} \langle C, X\rangle + \varepsilon. \tag{81}$$

В этом случае $\langle C, \widehat{X} \rangle$ является $\varepsilon$-аппроксимацией OT-расстояния и $\widehat{X}$ является аппроксимацией транспортного плана.

Введем ряд обозначений. Для конечномерного векторного пространства $E$ обозначим через $E^*$ сопряженное к нему пространство линейных функций $\langle g, x \rangle$, $x \in E$, $g \in E^*$. Обозначим через $\|\cdot\|_E$ выбранную норму на $E$ и через $\|\cdot\|_{E,*}$ норму на $E^*$, которая является сопряженной к $\|\cdot\|_E$. Для линейного оператора $A : E \to H$, определим его норму как $\|A\|_{E \to H} = \max_{x \in E, u \in H^*} \{\langle u, Ax \rangle : \|x\|_E = 1, \|u\|_{H,*} = 1\}$. Будем говорить, что функция $f : E \to \mathbb{R}$ является $\gamma$-сильно выпуклой на множестве $Q \subseteq E$ по отношению к норме на $E$, если для любых $x, y \in Q$, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\gamma}{2}\|x - y\|_E^2$, где $\nabla f(x)$ произвольный субградиент $f(x)$ в точке $x$.

Для матрицы $A$ и вектора $a$, обозначим $e^A$, $e^a$, $\ln A$, $\ln a$ покомпонентное применение соответствующих функций. Для вектора $a \in \mathbb{R}^n$, обозначим через $\|a\|_1$ сумму модулей его компонент, через $\|a\|_2$ его Евклидову норму, и через $\|a\|_\infty$ максимальный модуль его компонент. Для заданной матрицы $A \in \mathbb{R}^{n \times n}$ обозначим через $\mathrm{vec}(A)$ вектор из $\mathbb{R}^{n^2}$, получающийся из $A$ путем записи ее столбцов один над другим. Для матрицы $A \in \mathbb{R}^{n \times n}$ обозначим $\|A\|_1 = \|\mathrm{vec}(A)\|_1$ и $\|A\|_\infty = \|\mathrm{vec}(A)\|_\infty$. Также определим энтропию матрицы $X \in \mathbb{R}_+^{n \times n}$ как

$$H(X) := -\sum_{i,j=1}^{n} X^{ij} \ln X^{ij}. \tag{82}$$

Для двух матриц $A, B$ обозначим их Фробениусово скалярное произведение как $\langle A, B \rangle$. Обозначим через $\Delta^n := \{a \in \mathbb{R}_+^n : a^T \mathbf{1} = 1\}$ вероятностный симплекс в $\mathbb{R}^n$.

Для начала рассмотрим следующую общую задачу минимизации с линейными ограничениями

$$\min_{x \in Q \subseteq E} \{f(x) : Ax = b\}, \tag{83}$$

где $E$ – конечномерное векторное пространство, $Q$ – простое замкнутое выпуклое множество, $A$ – заданный линейный оператор из $E$ в некоторое конечномерное векторное пространство $H$, $b \in H$ задано, $f(x)$ – $\gamma$-сильно выпуклая функция на $Q$ по отношению к выбранной норме $\|\cdot\|_E$ на $E$. Двойственная по Лагранжу для задачи (83), записанная как задача минимизации, выглядит следующим образом

$$\min_{\lambda \in H^*} \left\{ \varphi(\lambda) := \langle \lambda, b \rangle + \max_{x \in Q} \left( -f(x) - \langle A^T \lambda, x \rangle \right) \right\}. \tag{84}$$

Заметим, что $\nabla \varphi(\lambda) = b - Ax(\lambda)$ непрерывен по Липшицу [36]

$$\|\nabla \varphi(\lambda_1) - \nabla \varphi(\lambda_1)\|_H \leq L \|\lambda_1 - \lambda_2\|_{H,*},$$

**Algorithm 12** Адаптивный прямо-двойственный ускоренный градиентный метод (АПДУГМ)

**Вход:** Точность $\varepsilon_f, \varepsilon_{eq} > 0$, начальная оценка $L_0$ такая, что $0 < L_0 < 2L$.

1: Положить $i_0 = k = 0$, $M_{-1} = L_0$, $\beta_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.
2: **repeat** {Основная итерация}
3: **repeat** {Линейный поиск}
4: Положить $M_k = 2^{i_k - 1} M_k$, найти $\alpha_{k+1}$ такое, что $\beta_{k+1} := \beta_k + \alpha_{k+1} = M_k \alpha_{k+1}^2$. Положить $\tau_k = \alpha_{k+1}/\beta_{k+1}$.
5: $\lambda_{k+1} = \tau_k \zeta_k + (1 - \tau_k)\eta_k$.
6: $\zeta_{k+1} = \zeta_k - \alpha_{k+1} \nabla \varphi(\lambda_{k+1})$.
7: $\eta_{k+1} = \tau_k \zeta_{k+1} + (1 - \tau_k)\eta_k$.
8: **until**

$$\varphi(\eta_{k+1}) \leq \varphi(\lambda_{k+1}) + \langle \nabla \varphi(\lambda_{k+1}), \eta_{k+1} - \lambda_{k+1} \rangle + \frac{M_k}{2} \|\eta_{k+1} - \lambda_{k+1}\|_2^2.$$

9: $\hat{x}_{k+1} = \tau_k x(\lambda_{k+1}) + (1 - \tau_k)\hat{x}_k$.
10: Положить $i_{k+1} = 0$, $k = k + 1$.
11: **until** $f(\hat{x}_{k+1}) + \varphi(\eta_{k+1}) \leq \varepsilon_f$, $\|A\hat{x}_{k+1} - b\|_2 \leq \varepsilon_{eq}$.
**Выход:** $\hat{x}_{k+1}$, $\eta_{k+1}$.

где $x(\lambda) := \arg\min_{x \in Q} \left( -f(x) - \langle A^T \lambda, x \rangle \right)$ и $L \leq \frac{\|A\|_{E \to H}^2}{\gamma}$. Эта оценка может быть завышенной и предлагаемый алгоритм не использует эту оценку и автоматически адаптируется к локальной константе Липшица градиента.

Предположим, что двойственная задача (84) имеет решение и существует такое $R > 0$, что $\|\lambda^*\|_2 \leq R < +\infty$, где $\lambda^*$ является решением (84) с минимальным значением $\|\lambda^*\|_2$.

**Теорема 3.3.** *Предположим, что целевая функция в прямой задаче* (83) *является $\gamma$-сильно выпуклой и что двойственное решение $\lambda^*$ удовлетворяет $\|\lambda^*\|_2 \leq R$. Тогда для $k \geq 1$ точки $\hat{x}_k$, $\eta_k$ в Алгоритме 12 удовлетворяют*

$$f(\hat{x}_k) - f^* \leq f(\hat{x}_k) + \varphi(\eta_k) \leq \frac{16 \|A\|_{E \to H}^2 R^2}{\gamma k^2}, \tag{85}$$

$$\|A\hat{x}_k - b\|_2 \leq \frac{16 \|A\|_{E \to H}^2 R}{\gamma k^2}, \quad \|\hat{x}_k - x^*\|_E \leq \frac{8}{k} \frac{\|A\|_{E \to H} R}{\gamma}, \tag{86}$$

*где $x^*$ и $f^*$ соответственно оптимальное решение и оптимальное значение функции в* (83). *Кроме того, критерий остановки на шаге 11 корректно определен.*

**Algorithm 13** Аппроксимация ОТ с помощью АПДУГМ

**Вход:** Точность $\varepsilon$.

Положить $\gamma = \frac{\varepsilon}{3 \ln n}$.
**for** $k = 1, 2, ...$ **do**
  Сделать шаг АПДУГМ и вычислить $\widehat{X}_k$ и $\eta_k$.
  Найти $\widehat{X}$ как проекцию $\widehat{X}_k$ на $\mathcal{U}(r, c)$ с помощью Алгоритма 2 из [37].
  **if** $\langle C, \widehat{X} - \widehat{X}_k \rangle \leq \frac{\varepsilon}{6}$ и $f(\hat{x}_k) + \varphi(\eta_k) \leq \frac{\varepsilon}{6}$ **then**
    Выдать $\widehat{X}$.
  **else**
    Положить $k = k + 1$ и продложить.
  **end if**
**end for**

Теперь применим общий метод, чтобы получить оценку сложности поиска транспортного плана $\widehat{X} \in \mathcal{U}(r, c)$, удовлетворяющего (81). Будем использовать энтропийную регуляризацию задачи (79) и рассмотрим регуляризованную задачу (80) с регуляризатором $\mathcal{R}(X) = -H(X)$, где $H(X)$ задано в (82). Определим $E = \mathbb{R}^{n^2}$, $\|\cdot\|_E = \|\cdot\|_1$, переменную $x = \text{vec}(X) \in \mathbb{R}^{n^2}$ соответствующую матрице $X$. Также положим $f(x) = \langle C, X \rangle - \gamma H(X)$, $Q = \Delta^{n^2}$, $b^T = (r^T, c^T)$ и $A : \mathbb{R}^{n^2} \to \mathbb{R}^{2n}$, исходя из соотношения $(A \,\text{vec}(X))^T = ((X\mathbf{1})^T, (X^T\mathbf{1})^T)$. Используя эти объекты, будем решать задачу (83) с помощью АПДУГМ. Пусть $\widehat{X}_k$ определяется соотношением $\text{vec}(\widehat{X}_k) = \hat{x}_k$, где $\hat{x}_k$ сгенерирован АПДУГМ. Также определим $\widehat{X} \in \mathcal{U}(r, c)$ как проекцию $\widehat{X}_k$ на $\mathcal{U}(r, c)$, полученную с помощью Алгоритма 2 в [37]. Псевдокод алгоритма для аппроксимации ОТ-расстояния приведен как Алгоритм 13.

**Теорема 3.4.** *Алгоритм 13 находит транспортный план $\widehat{X} \in \mathcal{U}(r, c)$, удовлетворяющий* (81) *за*

$$O\left(\min\left\{\frac{n^{9/4}\sqrt{R\|C\|_\infty \ln n}}{\varepsilon}, \frac{n^2 R\|C\|_\infty \ln n}{\varepsilon^2}\right\}\right) \tag{87}$$

*арифметических операций.*

### 3.3 Распределенный прямо-двойственный ускоренный стохастический градиентный метод

Результаты этого раздела опубликованы в статье [38].

Начнем с некоторых обозначений. Пусть $\mathcal{M}_+^1(\mathcal{X})$ множество положительных Радоновых вероятностных мер на метрическом пространстве $\mathcal{X}$, $S_1(n) = \{a \in \mathbb{R}_+^n \mid \sum_{l=1}^n a_l = 1\}$ – вероятностный симплекс. Обозначим через $\mathcal{C}(\mathcal{X})$ пространство непрерывных функций на $\mathcal{X}$, через $\delta(x)$

Дираковскую массу в точке $x$. Через $\lambda_{\max}(W)$ обозначим максимальное собственное число матрицы $W$. Также будем использовать жирный прямой шрифт для обозначения векторов $\mathtt{p} = [p_1^T, \cdots, p_m^T]^T \in \mathbb{R}^{mn}$, составленных из векторов $p_1, ..., p_m \in \mathbb{R}^n$. В этом случае $[\mathtt{p}]_i = p_i$ – $i$-й блок $\mathtt{p}$. Для вектора $\lambda \in \mathbb{R}^n$, обозначим через $[\lambda]_l$ его $l$-ю компоненту. Будем называть 2-нормой Евклидову норму $\|p\|_2 := \sqrt{\sum_{l=1}^n ([p]_l)^2}$.

Следуя ряду работ, начавшемуся с [39], рассмотрим энтропийную регуляризацию задачи оптимального транспорта (OT). Предположим, что задана положительная Радонова вероятностная мера $\mu$ с плотностью вероятности $q(y)$ на метрическом пространстве $\mathcal{Y}$ и дискретная вероятностная мера $\nu = \sum_{i=1}^n p_i \delta(z_i)$ с весами $p$ и конечным носителем в точках $z_1, \ldots, z_n \in \mathcal{Z}$ из метрического пространства $\mathcal{Z}$. Регуляризованное расстояние Васерштейна между $\mu$ и $\nu$ в полудискретной постановке определяется как

$$\mathcal{W}_\gamma(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \left\{ \sum_{i=1}^n \int_{\mathcal{Y}} c_i(y)\pi_i(y)dy + \gamma KL(\pi|\xi) \right\},$$

где $c_i(y) = c(z_i, y)$ – функция затрат на транспортировку единичной массы из точки $z_i$ в точку $y$, $\xi$ – равномерное распределение на $\mathcal{Y} \times \mathcal{Z}$, $KL(\pi|\xi) = \sum_{i=1}^n \int_{\mathcal{Y}} \pi_i(y) \log\left(\frac{\pi_i(y)}{\xi}\right) dy$, и множество допустимых транспортных планов $\pi$ определяется как

$$\Pi(\mu, \nu) = \left\{ \pi \in \mathcal{M}_+^1(\mathcal{Y}) \times S_1(n) : \sum_{i=1}^n \pi_i(y) = q(y), y \in \mathcal{Y}, \int_{\mathcal{Y}} \pi_i(y)dy = p_i, \forall\ i = 1, \ldots, n \right\}.$$

Для набора положительных Радоновых вероятностных мер $(\mu_1, \ldots, \mu_m)$ регуляризованный барицентр Васерштейна в полудискретной постановке определяется как решение $p$ задачи

$$\min_{p \in S_1(n)} \sum_{i=1}^m \mathcal{W}_{\gamma,\mu_i}(p) = \min_{\substack{p_1 = \cdots = p_m \\ p_1, \ldots, p_m \in S_1(n)}} \sum_{i=1}^m \mathcal{W}_{\gamma,\mu_i}(p_i), \tag{88}$$

где носитель $z_1, \ldots, z_n \in \mathcal{Z}$ барицентра $\nu$ зафиксирован и барицентр задается вектором $p \in S_n(1)$ таким, что $\nu = \sum_{i=1}^n p_i \delta(z_i)$. Также для краткости $\mathcal{W}_{\gamma,\mu}(p) := \mathcal{W}_\gamma(\mu, \nu)$. Целью раздела является разработка алгоритма для вычисления барицентра в постановке распределенной оптимизации.

Опишем теперь постановку задачи распределенной оптимизации для решения второй задачи в (88). Предположим, что есть сеть агентов, например, вычислительных машин, и каждая мера $\mu_i$ соответствует агенту $i$ и ему доступна возможность делать выборку случайной величины из этой меры. Математически сеть представлена связным ненаправленным графом $\mathcal{G} = (V, E)$, где $V$ – множество из $m$ вершин и $E$ – множество

ребер. Предполагается, что в графе нет вырожденных циклов. Структура сети приводит к информационным ограничениям, а именно, каждый узел $i$ имеет доступ только к $\mu_i$, а также может обмениваться информацией только с непосредственными соседями, т.е. узел $i$ может обмениваться информацией с узлом $j$ если и только если $(i,j) \in E$.

Представим ограничения на коммуникацию в сети с помощью одного ограничения вместо набора ограничений $p_1 = \cdots = p_m$ в (88). Для этого определим матрицу Лапласа $\bar{W} \in \mathbb{R}^{m\times m}$ для графа $\mathcal{G}$ как a) $[\bar{W}]_{ij} = -1$ если $(i,j) \in E$, b) $[\bar{W}]_{ij} = \deg(i)$ если $i = j$, c) $[\bar{W}]_{ij} = 0$ во всех остальных случаях. Здесь $\deg(i)$ обозначает число соседей узла $i$. Определим также матрицу коммуникаций как $W := \bar{W} \otimes I_n$.

Получаем, что $\sqrt{W}\mathtt{p} = 0$ тогда и только тогда, когда $p_1 = \cdots = p_m$. Используя этот факт, эквивалентно перепишем задачу (88) как задачу максимизации с линейными ограничениями

$$\max_{\substack{p_1,\ldots,p_m \in S_1(n) \\ \sqrt{W}\mathtt{p}=0}} -\sum_{i=1}^{m} \mathcal{W}_{\gamma,\mu_i}(p_i). \tag{89}$$

Введем вектор двойственных переменных $\boldsymbol{\lambda} = [\lambda_1^T, \cdots, \lambda_m^T]^T \in \mathbb{R}^{mn}$ для линейных ограничений $\sqrt{W}\mathtt{p} = 0$ в (89). Тогда двойственная по Лагранжу задача для (89) имеет вид

$$\begin{aligned} &\min_{\boldsymbol{\lambda}\in\mathbb{R}^{mn}} \max_{p_1,\ldots,p_m \in S_1(n)} \left\{ \sum_{i=1}^{m} \langle \lambda_i, [\sqrt{W}\mathtt{p}]_i \rangle - \mathcal{W}_{\gamma,\mu_i}(p_i) \right\} \\ &= \min_{\boldsymbol{\lambda}\in\mathbb{R}^{mn}} \sum_{i=1}^{m} \mathcal{W}^*_{\gamma,\mu_i}([\sqrt{W}\boldsymbol{\lambda}]_i), \end{aligned} \tag{90}$$

где $[\sqrt{W}\mathtt{p}]_i$ и $[\sqrt{W}\boldsymbol{\lambda}]_i$ обозначают $i$-й $n$-мерный блок векторов $\sqrt{W}\mathtt{p}$ и $\sqrt{W}\boldsymbol{\lambda}$ соответственно, и $\mathcal{W}^*_{\gamma,\mu_i}(\cdot)$ – сопряженная по Фенхелю для $\mathcal{W}_{\gamma,\mu_i}(p_i)$.

Далее рассмотрим общую задачу гладкой стохастической оптимизации, которая является двойственной для некоторой задачи с линейными ограничениями-равенствами. Задача (90) является частным случаем этой общей постановки. Для некоторого конечномерного векторного пространства $E$ обозначим через $E^*$ его двойственное. Пусть $\|\cdot\|_E$ обозначает некоторую норму на $E$ и $\|\cdot\|_{E,*}$ обозначает норму на $E^*$ двойственную к $\|\cdot\|_E$: $\|\lambda\|_{E,*} = \max_{\|x\|_E \le 1} \langle \lambda, x \rangle$. Для линейного оператора $A : E_1 \to E_2$, определим сопряженный оператор $A^T : E_2^* \to E_1^*$ как $\langle u, Ax \rangle = \langle A^T u, x \rangle, \quad \forall u \in E_2^*, \quad x \in E_1$. Будем говорить, что функция $f : E \to \mathbb{R}$ имеет $L$-Липшицев градиент в норме $\|\cdot\|_{E,*}$, если она дифференцируема и ее градиент удовлетворяет условию Липшица $\|\nabla f(x) - \nabla f(y)\|_{E,*} \le L\|x - y\|_E, \quad \forall x, y \in E$.

Для начала предложим прямо-двойственный алгоритм для общей

прямо-двойственной пары задач

$$(P) \qquad \min_{x \in Q \subseteq E} \{f(x) : Ax = b\}, \quad (D) \quad \min_{\lambda \in \Lambda} \left\{ \langle \lambda, b \rangle + \max_{x \in Q} \left( -f(x) - \langle A^T \lambda, x \rangle \right) \right\}.$$

где $Q$ – простое замкнутое выпуклое множество, $A : E \to H$ – заданный линейный оператор, $b \in H$ дано, $\Lambda = H^*$. Определим

$$\varphi(\lambda) := \langle \lambda, b \rangle + \max_{x \in Q} \left( -f(x) - \langle A^T \lambda, x \rangle \right) = \langle \lambda, b \rangle + f^*(-A^T \lambda) \tag{91}$$

и предположим, что это функция с $L$-Липшицевым градиентом. Здесь $f^*$ – сопряженная по Фенхелю для $f$. Предположим также, что $f^*(-A^T\lambda) = \mathbb{E}_\xi F^*(-A^T\lambda, \xi)$, где $\xi$ – случайный вектор. Также определим $F(x, \xi)$ как сопряженную по Фенхелю для $F^*$, т.е. удовлетворяющую $F^*(-A^T\lambda, \xi) = \max_{x \in Q}\{\langle -A^T\lambda, x\rangle - F(x, \xi)\}$, и $x(\lambda, \xi)$ как решение этой задачи максимизации. При этих предположениях, в двойственной задаче $(D)$ доступен стохастический оракул $(\Phi(x, \xi), \nabla\Phi(\lambda, \xi)) = (F^*(-A^T\lambda, \xi), \nabla F^*(-A^T\lambda, \xi))$, удовлетворяющий $\mathbb{E}_\xi \Phi(\lambda, \xi) = \varphi(\lambda)$, $\mathbb{E}_\xi \nabla\Phi(\lambda, \xi) = \nabla\varphi(\lambda)$, и который используется в нашем алгоритме. Наконец, предположим, что двойственная задача $(D)$ имеет решение $\lambda^*$ и существует некоторое $R > 0$ такое, что $\|\lambda^*\|_2 \leq R < +\infty$.

Дополнительно предполижим, что дисперсия стохастической аппроксимации $\nabla\Phi(\lambda, \xi)$ для градиента $\varphi$ может быть сделана сколь угодно маленькой, например, с использованием мини-батча. Также, так как $\nabla\Phi(\lambda, \xi) = b - A\nabla F^*(-A^T\lambda, \xi) = b - Ax(\lambda, \xi)$, то на каждой итерации, чтобы найти $\nabla\Phi(\lambda, \xi)$ сначала находится вектор $x(\lambda, \xi)$, который используется для итераций в прямом пространстве и аппроксимации решения прямой задачи.

**Algorithm 14** Ускоренный прямо-двойственный стохастический градиентный метод (УПДСГМ)

**Вход:** Число итераций $N$.

1: $C_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.
2: **for** $k = 0, \ldots, N-1$ **do**
3: Найти $\alpha_{k+1}$ как наибольший положительный корень уравнения $C_{k+1} := C_k + \alpha_{k+1} = 2L\alpha_{k+1}^2$. $\tau_{k+1} = \alpha_{k+1}/C_{k+1}$.
4: $\lambda_{k+1} = \tau_{k+1}\zeta_k + (1 - \tau_{k+1})\eta_k$
5: $\zeta_{k+1} = \zeta_k - \alpha_{k+1}\nabla\Phi(\lambda_{k+1}, \xi_{k+1})$.
6: $\eta_{k+1} = \tau_{k+1}\zeta_{k+1} + (1 - \tau_{k+1})\eta_k$.
7: $\hat{x}_{k+1} = \tau_{k+1}x(\lambda_{k+1}, \xi_{k+1}) + (1 - \tau_{k+1})\hat{x}_k$.
8: **end for**

**Выход:** Точки $\hat{x}_{k+1}$, $\eta_{k+1}$.

**Теорема 3.5.** *Пусть $\varphi$ обладает $L$-Липшицевым градиентом в 2-норме и $\|\lambda^*\|_2 \leq R$, где $\lambda^*$ – решение двойственной задачи $(D)$. Для заданной*

*точности решения $\varepsilon$, предположим, что на каждой итерации Алгоритма 14, стохастический градиент $\nabla\Phi(\lambda_k, \xi_k)$ выбран таким образом, что $\mathbb{E}_\xi\|\nabla\Phi(\lambda_k,\xi_k)-\nabla\varphi(\lambda_k)\|_2^2 \le \frac{\varepsilon L\alpha_k}{C_k}$. Тогда, для любого $\varepsilon > 0$ и $N \ge 0$, а также матожидания $\mathbb{E}$ по отношению ко всей случайности $\xi_1, \dots, \xi_N$, выход $\eta_N$ и $\hat{x}_N$ алгоритма 14 удовлетворяет*

$$f(\mathbb{E}\hat{x}_N) - f^* \le \frac{32LR^2}{N^2} + \frac{\varepsilon}{2} \quad и \quad \|A\mathbb{E}\hat{x}_N - b\|_2 \le \frac{32LR}{N^2} + \frac{\varepsilon}{2R}, \tag{92}$$

Применим общий алгоритм, описанный выше, для решения прямо-двойственной пары задач (89)-(90) и аппроксимации регуляризованного барицентра Васерштейна, являющегося решением задачи (89). Начнем со вспомогательной леммы.

**Лемма 3.2.** *Градиент целевой функции $\mathcal{W}_\gamma^*(\boldsymbol{\lambda})$ в двойственной задаче (90) является $\lambda_{\max}(W)/\gamma$-Липшицевым в 2-норме. Пусть стохастическая аппроксимация этого градиента определена как*

$$[\widetilde{\nabla}\mathcal{W}_\gamma^*(\boldsymbol{\lambda})]_i = \sum_{j=1}^m \sqrt{W}_{ij}\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_j}(\bar{\lambda}_j),\ i = 1, ..., m,\ c$$

$$\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_j}(\bar{\lambda}_j) = \frac{1}{M}\sum_{r=1}^M p_j(\bar{\lambda}_j),\ u\ [p_j(\bar{\lambda}_j)]_l = \frac{\exp(([\bar{\lambda}_j]_l - c_l(Y_r^j))/\gamma)}{\sum_{\ell=1}^n \exp(([\bar{\lambda}_j]_\ell - c_\ell(Y_r^j))/\gamma)}. \tag{93}$$

*где $M$ – размер мини-батча, $\bar{\lambda}_j := [\sqrt{W}\boldsymbol{\lambda}]_j$, $j = 1, ..., m$, $Y_1^j, ..., Y_r^j$ – выборка из меры $\mu_j$, $j = 1, ..., m$. Тогда $\mathbb{E}_{Y_r^j\sim\mu_j, j=1,...,m, r=1,...,M}\widetilde{\nabla}\mathcal{W}_\gamma^*(\boldsymbol{\lambda}) = \nabla\mathcal{W}_\gamma^*(\boldsymbol{\lambda})$ и*

$$\mathbb{E}_{Y_r^j\sim\mu_j, j=1,...,m, r=1,...,M}\|\widetilde{\nabla}\mathcal{W}_\gamma^*(\boldsymbol{\lambda}) - \nabla\mathcal{W}_\gamma^*(\boldsymbol{\lambda})\|_2^2 \le \frac{\lambda_{\max}(W)}{M},\ \boldsymbol{\lambda} \in \mathbb{R}^{mn}. \tag{94}$$

Из этой леммы следует, что если на каждой итерации Алгоритма 14 размер мини-батча $M_k$ удовлетворяет $M_k \ge \frac{\lambda_{\max}(W)C_k}{L\alpha_k\varepsilon}$, то выполнены предположения Теоремы 3.5.

Для конкретной задачи (90) шаг 5 Алгоритма 14 может быть записан поблочно $[\boldsymbol{\zeta}_{k+1}]_i = [\boldsymbol{\zeta}_k]_i - \alpha_{k+1}\sum_{j=1}^m \sqrt{W}_{ij}\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_j}([\sqrt{W}\boldsymbol{\lambda}_{k+1}]_j)$, $i = 1, ..., m$. Сделаем замену переменной и обозначим $\bar{\boldsymbol{\lambda}} = \sqrt{W}\boldsymbol{\lambda}$, $\bar{\boldsymbol{\eta}} = \sqrt{W}\boldsymbol{\eta}$, $\bar{\boldsymbol{\zeta}} = \sqrt{W}\boldsymbol{\zeta}$. Тогда шаг 5 Алгоритма 14 будет выглядеть как $[\bar{\boldsymbol{\zeta}}_{k+1}]_i = [\bar{\boldsymbol{\zeta}}_k]_i - \alpha_{k+1}\sum_{j=1}^m W_{ij}\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_j}([\bar{\boldsymbol{\lambda}}_{k+1}]_j)$, $i = 1, ..., m$. В отличие от исходного шага, этот шаг может быть выполнен с использованием только локальных коммуникаций между соседями в вычислительной сети.

**Algorithm 15** Распределенный алгоритм для вычисления барицентра Васерштейна

**Вход:** Каждому агенту $i \in V$ приписывается мера $\mu_i$.

1: Все агенты полагают $[\bar{\boldsymbol{\eta}}_0]_i = [\bar{\boldsymbol{\zeta}}_0]_i = [\bar{\boldsymbol{\lambda}}_0]_i = \mathbf{0} \in \mathbb{R}^n$, $C_0 = \alpha_0 = 0$ и $N$

2: Для каждого агента $i \in V$:

3: **for** $k = 0, \ldots, N-1$ **do**

4: Найти $\alpha_{k+1}$ как наибольший корень уравнения $C_{k+1} := C_k + \alpha_{k+1} = 2L\alpha_{k+1}^2$. $\tau_{k+1} = \alpha_{k+1}/C_{k+1}$.

5: Положить $M_{k+1} = \max\{1,\ \lambda_{\max}(W)C_{k+1}/(L\alpha_{k+1}\varepsilon)\}$

6: $[\bar{\boldsymbol{\lambda}}_{k+1}]_i = \tau_{k+1}[\bar{\boldsymbol{\zeta}}_k]_i + (1-\tau_{k+1})[\bar{\boldsymbol{\eta}}_k]_i$

7: Сгенерировать выборку $\{Y_r^i\}_{r=1}^{M_{k+1}}$ размера $M_{k+1}$ из меры $\mu_i$ и положить $\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_i}([\bar{\boldsymbol{\lambda}}_{k+1}]_i)$ в соответствии с (93).

8: Передать $\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_i}([\bar{\boldsymbol{\lambda}}_{k+1}]_i)$ соседям $\{j \mid (i,j) \in E\}$

9: $[\bar{\boldsymbol{\zeta}}_{k+1}]_i = [\bar{\boldsymbol{\zeta}}_k]_i - \alpha_{k+1}\sum_{j=1}^m W_{ij}\widetilde{\nabla}\mathcal{W}^*_{\gamma,\mu_j}([\bar{\boldsymbol{\lambda}}_{k+1}]_j)$

10: $[\bar{\boldsymbol{\eta}}_{k+1}]_i = \tau_{k+1}[\bar{\boldsymbol{\zeta}}_{k+1}]_i + (1-\tau_{k+1})[\bar{\boldsymbol{\eta}}_{k+1}]_i$

11: $[\hat{\mathtt{p}}_{k+1}]_i = \tau_{k+1}p_i([\bar{\boldsymbol{\lambda}}_{k+1}]_i) + (1-\tau_{k+1})[\hat{\mathtt{p}}_{k+1}]_i$, где $p_i(\cdot)$ определено в (93).

12: **end for**

**Выход:** $\hat{\mathtt{p}}_N$.

**Теорема 3.6.** *При перечисленных предположениях Алгоритм 15 после* $N = \sqrt{16\lambda_{max}(W)R^2/(\varepsilon\gamma)}$ *итераций возвращает аппроксимацию* $\hat{\mathtt{p}}_N$ *для барицентра, которая удовлетворяет*

$$\sum_{i=1}^m \mathcal{W}_{\gamma,\mu_i}(\mathbb{E}[\hat{\mathtt{p}}_N]_i) - \sum_{i=1}^m \mathcal{W}_{\gamma,\mu_i}([\mathtt{p}^*]_i) \leq \varepsilon, \quad \|\sqrt{W}\mathbb{E}\hat{\mathtt{p}}_N\|_2 \leq \varepsilon/R. \tag{95}$$

*Кроме того, общая сложность вычислений составляет*

$$O\left(n\max\lambda_{\max}(W)R^2/\varepsilon^2, \sqrt{\lambda_{\max}(W)R^2/(\varepsilon\gamma)}\right)$$

*арифметических операций.*

## 3.4 Прямо-двойственный ускоренный градиентный метод с малоразмерной минимизацией

Результаты этого раздела опубликованы в статьях [40, 41].

Рассмотрим следующую задачу

$$(P_1) \qquad \min_{x \in Q \subseteq E}\{f(x) : \mathbf{A}x = b\},$$

где $E$ – конечномерное векторное пространство, $Q$ – простое замкнутое выпуклое множество, $\mathbf{A}$ – заданный линейный оператор из $E$ в некоторое

конечномерное векторное пространство $H$, $b \in H$ задано. Двойственной по Лагранжу для этой задачи является

$$(D_1) \qquad \max_{\lambda \in \Lambda} \left\{ -\langle \lambda, b \rangle + \min_{x \in Q} \left( f(x) + \langle \mathbf{A}^T \lambda, x \rangle \right) \right\}.$$

Здесь мы обозначили $\Lambda = H^*$. Перепишем задачу $(D_1)$ эквивалентно как задачу минимизации

$$(P_2) \quad \min_{\lambda \in \Lambda} \left\{ \langle \lambda, b \rangle + \max_{x \in Q} \left( -f(x) - \langle \mathbf{A}^T \lambda, x \rangle \right) \right\}.$$

Обозначим

$$\varphi(\lambda) = \langle \lambda, b \rangle + \max_{x \in Q} \left( -f(x) - \langle \mathbf{A}^T \lambda, x \rangle \right). \tag{96}$$

Так как $f$ выпуклая, то $\varphi(\lambda)$ также выпуклая и по теореме Демьянова-Данскина ее субградиент равен (см., например [36]) $\nabla \varphi(\lambda) = b - \mathbf{A}x(\lambda)$ где $x(\lambda)$ – некоторое решение задачи $\max_{x \in Q} \left( -f(x) - \langle \mathbf{A}^T \lambda, x \rangle \right)$.

Сделаем следующие предположения о двойственной задаче $(D_1)$

- Субградиент целевой функции $\varphi(\lambda)$ удовлетворяет условию Гельдера с костантой $M_\nu$, т.е., для любых $\lambda, \mu \in \Lambda$ и некоторого $\nu \in [0, 1]$ выполнено $\|\nabla \varphi(\lambda) - \nabla \varphi(\mu)\|_* \leqslant M_\nu \|\lambda - \mu\|^\nu$.
- Двойственная задача $(D_1)$ имеет решение $\lambda^*$ и существует число $R > 0$ такое, что $\|\lambda^*\|_2 \leqslant R < +\infty$.

Выберем Евклидову норму и соответствующую прокс-функцию $d(\lambda) = \frac{1}{2}\|\lambda\|_2^2$ в двойственном пространстве векторов $\lambda$. Тогда соответствующая дивергенция Брегмана равна $V[\zeta](\lambda) = \frac{1}{2}\|\lambda - \zeta\|_2^2$. Предложенный прямо-двойственный метод для задачи $(P_1)$ приведен ниже как Алгоритм 16.

**Теорема 3.7.** *Пусть целевая функция $\varphi$ в задаче $(P_2)$ имеет Гельдеров субградиент, а также ограниченное решение, т.е. $\|\lambda^*\|_2 \leqslant R$. Тогда для последовательности $\hat{x}^{k+1}, \eta^{k+1}$, $k \geqslant 0$, сгенерированной Алгоритмом 16 справедливо*

$$\|\mathbf{A}\hat{x}^k - b\|_2 \leqslant \frac{2R}{A_k} + \frac{\varepsilon}{2R}, \quad |\varphi(\eta^k) + f(\hat{x}^k)| \leqslant \frac{2R^2}{A_k} + \frac{\varepsilon}{2}, \tag{97}$$

*где* $A_k \geqslant \left[\frac{1+\nu}{1-\nu}\right]^{\frac{1-\nu}{1+\nu}} \frac{k^{\frac{1+3\nu}{1+\nu}} \varepsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{1+3\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}}$.

Сделаем замечание относительно соответствующих оценок сложности. Как следует из Теоремы 3.7, если $A_k \geqslant 2R^2/\varepsilon$, то невязка по целевой функции и по ограничениям становится меньше $\varepsilon$. В то же время, используя нижнюю оценку для $A_k$, получим, что число итераций для

**Algorithm 16** Прямо-двойственный универсальный градиентный метод с маломерной минимизацией

**Вход:** Начальная точка $\lambda_0 = 0$, точность $\tilde{\varepsilon}_f, \tilde{\varepsilon}_{eq} > 0$.

1: Положить $k = 0$, $A_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.
2: **repeat**
3: $\beta_k = \operatorname{argmin}_{\beta \in [0,1]} \varphi\left(\zeta^k + \beta(\eta^k - \zeta^k)\right)$; $\lambda^k = \zeta^k + \beta_k(\eta^k - \zeta^k)$
4: $h_{k+1} = \operatorname{argmin}_{h \geqslant 0} \varphi\left(\lambda^k - h\nabla\varphi(\lambda^k)\right)$; $\eta^{k+1} = \lambda^k - h_{k+1}\nabla\varphi(\lambda^k)$ // Выбрать $\nabla\varphi(\lambda^k) : \langle \nabla\varphi(\lambda^k), \zeta^k - \lambda^k \rangle \geqslant 0$
5: Найти $a_{k+1}$ из уравнения $\varphi(\eta^{k+1}) = \varphi(\lambda^k) - \frac{a_{k+1}^2}{2A_{k+1}}\|\nabla\varphi(\lambda^k)\|_2^2 + \frac{\varepsilon a_{k+1}}{2A_{k+1}}$ // $A_{k+1} = A_k + a_{k+1}$
6: $\zeta^{k+1} = \zeta^k - a_{k+1}\nabla\varphi(\lambda^k)$
7: Положить

$$\hat{x}^{k+1} = \frac{1}{A_{k+1}} \sum_{i=0}^{k} a_{i+1} x(\lambda^i) = \frac{a_{k+1}x(\lambda^k) + A_k\hat{x}^k}{A_{k+1}}.$$

8: Положить $k = k + 1$.
9: **until** $|f(\hat{x}^{k+1}) + \varphi(\eta^{k+1})| \leqslant \tilde{\varepsilon}_f$, $\|\mathbf{A}\hat{x}^{k+1} - b\|_2 \leqslant \tilde{\varepsilon}_{eq}$.

**Выход:** Точки $\hat{x}^{k+1}$, $\eta^{k+1}$.

достижения точности $\varepsilon$ равно $O\left(\left(\frac{M_\nu^{\frac{2}{1+\nu}} R^2}{\varepsilon^{\frac{2}{1+\nu}}}\right)^{\frac{1+\nu}{1+3\nu}}\right)$. Так как алгоритм не использует значение параметра $\nu$, можно взять минимум от оценки сложности по $\nu \in [0, 1]$. Это означает, что предложенный алгоритм является равномерно оптимальным в классе задач с Гельдеровым градиентом, т.е. универсальным.

# 4 Заключение

В результате подготовки данной диссертации были опубликованы статьи [21, 22, 24, 27, 30, 33, 34, 38, 40, 41].

В статьях [21, 22, 24, 27] разработаны методы оптимизации со (стохастическим) неточным оракулом первого порядка, неточным оракулом нулевого порядка, оракулом неточной производной по направлению. Также рассмотрено конкретное приложение к задаче обучения модели ранжирования веб-страниц.

В статьях [30, 33, 34, 38, 40, 41] разработаны прямо-двойственные методы для выпуклых задач с линейными ограничениями. В частности, рассматриваются бесконечномерные седловые задачи, для которых получены оценки сложности, не зависящие от размерности. Также рассматриваются задачи (стохастической) выпуклой оптимизации с линейными ограничениями, для которых предложены ускоренные градиентные методы с оптимальной скоростью сходимости. Эти методы приме-

нены для приближенного решения задачи поиска ОТ расстояния и барицентра Васерштейна.

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

1. Стохастический промежуточный градиентный метод для выпуклых задач со стохастическим неточным оракулом.

2. Градиентный метод с неточным оракулом для детерминированной невыпуклой оптимизации и безградиентный метод с неточным оракулом для детерминированной выпуклой оптимизации.

3. Концепция неточного оракула для методов с производной по направлению, ускоренный и неускоренный метод с неточной производной по направлению для задач сильно выпуклой стохастической гладкой оптимизации.

4. Прямо-двойственные методы для бесконечномерных игр в выпукло-вогнутом и сильно выпукло-вогнутом случае.

5. Неадаптивный и адаптивный ускоренный прямо-двойственный градиентный метод для сильно выпуклых задач с линейными ограничениями равенствами и неравенствами.

6. Новые оценки сложности для задачи поиска оптимально-транспортого расстояния.

7. Стохастический прямо-двойственный ускоренный градиентный метод для задач с линейными ограничениями и его приложение к задаче аппроксимации барицентра Васерштейна.

8. Универсальный прямо-двойственный ускоренный градиентный метод с маломерной оптимизацией.

## Финансирование

Исследования, вошедшие в диссертацию, а также подготовка диссертации были поддержаны грантом Российского научного фонда (проект № 18-71-10108), грантами РФФИ 18-31-20005 мол-а-вед и 18-29-03071-мк, а также Министерством Науки и Высшего Образования Российской Федерации (Госзадание) №075-00337-20-03, проект № 0714-2020-0005.

## 5 Список литературы

[1] N. Karmarkar. «A new polynomial-time algorithm for linear programming». в: *Combinatorica* 4.4 (1984), с. 373—395. ISSN: 1439-6912. DOI: 10.1007/BF02579150.

[2] Yurii Nesterov и Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[3] Augustin Cauchy. «Méthode générale pour la résolution des systémes d'équations simultanées». в: *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 55 (1847), с. 536—538.

[4] Boris Polyak. «Gradient methods for the minimisation of functionals». в: *USSR Computational Mathematics and Mathematical Physics* 3.4 (1963), с. 864—878. ISSN: 0041-5553. DOI: http://dx.doi.org/10.1016/0041-5553(63)90382-3.

[5] Herbert Robbins и Sutton Monro. «A Stochastic Approximation Method». в: *Ann. Math. Statist.* 22.3 (сент. 1951), с. 400—407. DOI: 10.1214/aoms/1177729586.

[6] A.S. Nemirovsky и D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. J. Wiley & Sons, New York, 1983.

[7] Yurii Nesterov. «A method of solving a convex programming problem with convergence rate $O(1/k^2)$». в: *Soviet Mathematics Doklady* 27.2 (1983), с. 372—376.

[8] Amir Beck и Marc Teboulle. «A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems». в: *SIAM Journal on Imaging Sciences* 2.1 (2009), с. 183—202. DOI: 10.1137/080716542.

[9] Yurii Nesterov. «Gradient methods for minimizing composite functions». в: *Mathematical Programming* 140.1 (2013), с. 125—161.

[10] Guanghui Lan. «An optimal method for stochastic composite optimization». в: *Mathematical Programming* 133.1 (2012), с. 365—397. ISSN: 1436-4646. URL: https://doi.org/10.1007/s10107-010-0434-y.

[11] Rie Johnson и Tong Zhang. «Accelerating Stochastic Gradient Descent using Predictive Variance Reduction». в: *Advances in Neural Information Processing Systems 26*. под ред. C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani и K. Q. Weinberger. Curran Associates, Inc., 2013, с. 315—323. URL: http://papers.nips.cc/paper/4937-accelerating-stochastic-gradient-descent-using-predictive-variance-reduction.pdf.

[12] Qihang Lin, Zhaosong Lu и Lin Xiao. «An Accelerated Proximal Coordinate Gradient Method». в: *Advances in Neural Information Processing Systems 27*. под ред. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence и K. Q. Weinberger. First appeared in arXiv:1407.1296. Curran Associates, Inc., 2014, с. 3059—3067. URL: http://papers.nips.cc/paper/5356-an-accelerated-proximal-coordinate-gradient-method.pdf.

[13] Hongzhou Lin, Julien Mairal и Zaid Harchaoui. «A Universal Catalyst for First-order Optimization». в: *Proceedings of the 28th International Conference on Neural Information Processing Systems*. NIPS'15. Montreal, Canada: MIT Press, 2015, с. 3384—3392. URL: http://dl.acm.org/citation.cfm?id=2969442.2969617.

[14] Guanghui Lan и Yi Zhou. «An optimal randomized incremental gradient method». в: *Mathematical Programming* (2017). ISSN: 1436-4646. DOI: 10.1007/s10107-017-1173-0.

[15] Shai Shalev-Shwartz и Tong Zhang. «Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization». в: *Proceedings of the 31st International Conference on Machine Learning*. под ред. Eric P. Xing и Tony Jebara. т. 32. Proceedings of Machine Learning Research. First appeared in arXiv:1309.2375. Bejing, China: PMLR, 2014, с. 64—72. URL: http://proceedings.mlr.press/v32/shalev-shwartz14.html.

[16] Yurii Nesterov. «Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems». в: *SIAM Journal on Optimization* 22.2 (2012), с. 341—362. DOI: 10.1137/100802001.

[17] Yurii Nesterov и Vladimir Spokoiny. «Random Gradient-Free Minimization of Convex Functions». в: *Found. Comput. Math.* 17.2 (апр. 2017). First appeared in 2011 as CORE discussion paper 2011/16, с. 527—566. ISSN: 1615-3375. DOI: 10.1007/s10208-015-9296-2.

[18] Alexandre d'Aspremont. «Smooth Optimization with Approximate Gradient». в: *SIAM J. on Optimization* 19.3 (окт. 2008), с. 1171—1183. ISSN: 1052-6234. DOI: 10.1137/060676386.

[19] Olivier Devolder, François Glineur и Yurii Nesterov. «First-order methods of smooth convex optimization with inexact oracle». в: *Mathematical Programming* 146.1 (2014), с. 37—75. ISSN: 1436-4646. DOI: 10.1007/s10107-013-0677-5.

[20] Amir Beck и Marc Teboulle. «A fast dual proximal gradient algorithm for convex minimization and applications». в: *Operations Research Letters* 42.1 (2014), с. 1—6.

[21] Pavel Dvurechensky и Alexander Gasnikov. «Stochastic Intermediate Gradient Method for Convex Problems with Stochastic Inexact Oracle». в: *Journal of Optimization Theory and Applications* 171.1 (2016), с. 121—145.

[22] A. V. Gasnikov и P. E. Dvurechensky. «Stochastic intermediate gradient method for convex optimization problems». в: *Doklady Mathematics* 93.2 (2016), с. 148—151.

[23] Olivier Devolder. «Stochastic first order methods in smooth convex optimization». в: *CORE Discussion Paper 2011/70* (2011).

[24] Lev Bogolubsky, Pavel Dvurechensky, Alexander Gasnikov, Gleb Gusev, Yurii Nesterov, Andrei M Raigorodskii, Aleksey Tikhonov и Maksim Zhukovskii. «Learning Supervised PageRank with Gradient-Based and Gradient-Free Optimization Methods». в: *Advances in Neural Information Processing Systems 29*. под ред. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon и R. Garnett. Curran Associates, Inc., 2016, с. 4914—4922. URL: http://papers.nips.cc/paper/6565-learning-supervised-pagerank-with-gradient-based-and-gradient-free-optimization-methods.pdf.

[25] Yurii Nesterov и Arkadi Nemirovski. «Finding the stationary states of Markov chains by iterative methods». в: *Applied Mathematics and Computation* 255 (2015). Special issue devoted to the international conference "Numerical computations: Theory and Algorithms" June 17–23, 2013, Falerna, Italy, с. 58—65. ISSN: 0096-3003. DOI: https://doi.org/10.1016/j.amc.2014.04.053.

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

[27] Pavel Dvurechensky, Eduard Gorbunov и Alexander Gasnikov. «An Accelerated Directional Derivative Method for Smooth Stochastic Convex Optimization». в: *European Journal of Operational Research* (2020). ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2020.08.027.

[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] Anatoli Juditsky и Yuri Nesterov. «Deterministic and Stochastic Primal-Dual Subgradient Algorithms for Uniformly Convex Minimization». в: *Stochastic Systems* 4.1 (2014), с. 44—80. DOI: 10.1287/10-SSY010.

[30] Pavel Dvurechensky, Yurii Nesterov и Vladimir Spokoiny. «Primal-Dual Methods for Solving Infinite-Dimensional Games». в: *Journal of Optimization Theory and Applications* 166.1 (2015), с. 23—51.

[31] Yurii Nesterov. «Primal-dual subgradient methods for convex problems». в: *Mathematical Programming* 120.1 (2009), с. 221—259. ISSN: 1436-4646. DOI: 10.1007/s10107-007-0149-x.

[32] Yurii Nesterov. «Dual extrapolation and its applications to solving variational inequalities and related problems». в: *Mathematical Programming* 109.2-3 (2007). First appeared in 2003 as CORE discussion paper 2003/68, с. 319—344.

[33] Alexey Chernov, Pavel Dvurechensky и Alexander Gasnikov. «Fast Primal-Dual Gradient Method for Strongly Convex Minimization Problems with Linear Constraints». в: *Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19-23, 2016, Proceedings*. под ред. Yury Kochetov, Michael Khachay, Vladimir Beresnev, Evgeni Nurminski и Panos Pardalos. Springer International Publishing, 2016, с. 391—403.

[34] Pavel Dvurechensky, Alexander Gasnikov и Alexey Kroshnin. «Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm». в: *Proceedings of the 35th International Conference on Machine Learning*. под ред. Jennifer Dy и Andreas Krause. т. 80. Proceedings of Machine Learning Research. arXiv:1802.04367. 2018, с. 1367—1376.

[35] Leonid Kantorovich. «On the translocation of masses». в: *Doklady Acad. Sci. USSR (N.S.)* 37 (1942), с. 199—201.

[36] Yurii Nesterov. «Smooth minimization of non-smooth functions». в: *Mathematical Programming* 103.1 (2005), с. 127—152.

[37] Jason Altschuler, Jonathan Weed и Philippe Rigollet. «Near-linear time approxFimation algorithms for optimal transport via Sinkhorn iteration». в: *Advances in Neural Information Processing Systems 30*. под ред. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan и R. Garnett. arXiv:1705.09634. Curran Associates, Inc., 2017, с. 1961—1971.

[38] Pavel Dvurechensky, Darina Dvinskikh, Alexander Gasnikov, César A. Uribe и Angelia Nedić. «Decentralize and Randomize: Faster Algorithm for Wasserstein Barycenters». в: *Advances in Neural Information Processing Systems 31*. под ред. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi и R. Garnett. NeurIPS 2018. arXiv:1806.03915. Curran Associates, Inc., 2018, с. 10783—10793.

[39] Marco Cuturi. «Sinkhorn Distances: Lightspeed Computation of Optimal Transport». в: *Advances in Neural Information Processing Systems 26*. под ред. C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani и K. Q. Weinberger. Curran Associates, Inc., 2013, с. 2292—2300.

[40] S. V. Guminov, Yu. E. Nesterov, P. E. Dvurechensky и A. V. Gasnikov. «Accelerated Primal-Dual Gradient Descent with Linesearch for Convex, Nonconvex, and Nonsmooth Optimization Problems». в: *Doklady Mathematics* 99.2 (2019), с. 125—128.

[41] Yurii Nesterov, Alexander Gasnikov, Sergey Guminov и Pavel Dvurechensky. «Primal-dual accelerated gradient methods with small-dimensional relaxation oracle». в: *Optimization Methods and Software* (2020), с. 1—28. DOI: 10.1080/10556788.2020.1731747.