На правах рукописи



Газдиева Милена Джамбулатовна

Параметрические методы вычисления обобщенных оптимальных транспортных отображений и барицентров

Специальность 1.2.2. —
«Математическое моделирование, численные методы и комплексы программ»

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2026

Работа выполнена в Автономной некоммерческой образовательной организации высшего образования «Сколковский институт науки и технологий».

Научный руководитель: **Бурнаев Евгений Владимирович**
доктор физико-математических наук, профессор РАН, доцент, АНОО ВО «Сколковский институт науки и технологий», профессор, директор Центра искусственного интеллекта

Официальные оппоненты: **Орлов Юрий Николаевич**
доктор физико-математических наук, доцент, ФГУ «ФИЦ Институт прикладной математики им. М. В. Келдыша Российской академии наук», главный научный сотрудник, и. о. заведующего отделом №6

Турдаков Денис Юрьевич
кандидат физико-математических наук, ФГБУН Институт системного программирования им. В. П. Иванникова Российской академии наук, заведующий отделом «Информационные системы», руководитель Исследовательского центра доверенного искусственного интеллекта

Ведущая организация: ФГАОУ ВО «Московский физико-технический институт (национальный исследовательский университет)»

Защита состоится «23» декабря 2026 г. в 11 часов 00 минут на заседании диссертационного совета 24.1.224.01 на базе ФГУ «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН) по адресу: 117312, г. Москва, проспект 60-летия Октября, 9.

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН по адресу: 119333, г. Москва, ул. Вавилова, д. 44, корп. 2 и на официальном сайте: `https://www.frccsc.ru/`.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просьба направлять по адресу: 119333, г. Москва, ул. Вавилова, д. 44, корп. 2, ученому секретарю диссертационного совета 24.1.224.01.

Автореферат разослан «__» _____ 2026 года.
Телефон для справок: +7 (499) 135-51-64.

Ученый секретарь диссертационного
совета 24.1.224.01,
доктор техн. наук, доцент Смирнов Иван Валентинович

Основная характеристика работы

Актуальность темы исследования. В работе представлен ряд новых численных методов, относящихся к области генеративного моделирования (ГМ), цель которого заключается в построении математических моделей, описывающих закономерности данных и позволяющих синтезировать новые данные на их основе. Рассматривается специальный класс задач ГМ, в которых по заданному элементу из одного множества объектов требуется сгенерировать объект из другого множества, максимально сохраняющий выбранные характеристики исходного. Примерами таких задач являются сверхразрешение и перенос стиля изображений, где множества объектов обозначают классы изображений, а выбранные характеристики отражают содержание (контент) изображения. Таким образом, задача сводится к тому, чтобы для изображения одного класса создать (сгенерировать) изображение другого класса, максимально сохраняющее содержание исходного. В последние годы для решения задач ГМ, включая указанные задачи преобразования данных, стали активно применяться численные методы на основе математической теории оптимального транспорта (ОТ). Диссертация посвящена разработке новых методов ОТ, устраняющих ряд недостатков существующих методов и расширяющих область их применимости.

Оптимальный транспорт (ОТ) представляет собой совокупность математических задач, направленных на нахождение наиболее эффективного способа переноса массы между вероятностными распределениями, который называется ОТ отображением или планом в зависимости от постановки математической задачи. Основы данной теории были заложены в работах советского математика и экономиста Леонида Витальевича Канторовича в контексте решения прикладных задач поиска оптимального плана перевозки грузов и распределения ресурсов. Исторически большинство численных методов ОТ, таких как алгоритм Синкхорна, было ориентировано на работу с дискретными вероятностными распределениями, то есть распределениями с дискретным носителем. Такие методы, как правило, позволяют найти сопоставление между элементами конечных выборок, однако в задачах ГМ и преобразования данных требуется не только сопоставить элементы выборок, но и построить правило преобразования, применимое к новым объектам. Дискретные методы ОТ сами по себе не дают такого обобщаемого правила преобразования.

В контексте ГМ первые численные методы ОТ, среди которых WGAN и WGAN-GP, применялись преимущественно для вычисления функций потерь при обучении нейросетевых моделей. В последние годы стали активно развиваться методы ОТ, основанные на том, что многие задачи ГМ могут быть сформулированы как задачи ОТ между непрерывными распределениями. При этом предполагается, что распределения неизвестны и заданы выборками, по которым требуется построить решение задачи

ОТ. Такие методы, в том числе W2GN, NOT и Kernel NOT, используют выразительную параметризацию транспортных отображений, например, на основе глубоких нейросетей, и обычно называются *непрерывными* или *параметрическими* методами ОТ. В этом контексте вычисление ОТ рассматривается как обучение параметрической модели.

Непрерывные методы ОТ продемонстрировали определенные успехи в решении ряда практических задач ГМ, в частности, в задаче адаптации моделей к данным, поступающим из других источников. Данная задача важна в областях, где объёмы выборок ограничены и возникает необходимость объединения информации из различных источников, например в медицине, и может быть решена с использованием ОТ отображений или барицентров – специальных операций усреднения в пространстве распределений, применяемых для агрегации данных.

Однако, существующие непрерывные методы ОТ обладают значительными ограничениями, снижающими их эффективность в описанных выше задачах ГМ. В частности, такие методы чувствительны к неточностям в исходных данных, таким как *выбросы*, которые могут приводить к отклонениям вычисленных решений задачи ОТ от тех, которые были бы получены в условиях идеальных данных. Причина состоит в том, что указанные методы решают задачу ОТ в классической формулировке, накладывающей жесткое условие равенства на исходное и сгенерированное распределения, что может приводить к нежелательному переносу массы к аномальным точкам. Как следствие, данные методы не способны обеспечивать точные приближения решений задачи ОТ при работе с реальными данными, содержащими различного рода неточности.

Для преодоления этих ограничений все большее внимание уделяется методам, основанным на обобщенных формулировках ОТ, которые ослабляют упомянутое выше *условие равенства*, а именно на задачах частичного и несбалансированного ОТ. В задаче частичного ОТ требование равенства заменяется неравенством, что дозволяет частичный перенос массы между распределениями. В несбалансированном ОТ строгое условие сохранения массы заменяется регуляризационными штрафами, то есть рассматривается перенос массы между распределениями, которые близки к исходным, но не обязаны совпадать с ними. В результате практическая применимость непрерывных методов обобщенного ОТ существенно расширяется, поскольку они позволяют эффективно работать с данными, страдающими от различного рода неточностей.

Однако в области непрерывного обобщенного ОТ остается ряд нерешенных задач. Существующие методы несбалансированного ОТ можно разделить на две группы: минимаксные подходы, в том числе метод UOTM, использующие несколько нейросетей, и методы, основанные на использовании дискретных решений задачи несбалансированного ОТ при построении модели, среди которых – NubOT, GENOT. Минимаксные

методы, как правило, подкреплены теоретическими гарантиями сходимости, однако требуют решения задачи поиска седловой точки и настройки большого числа параметров нейросетей, что может приводить к неустойчивой и вычислительно затратной оптимизации. Методы второй группы не имеют такой проблемы, однако лишены теоретических гарантий.

Между тем параметрические методы частичного ОТ и методы решения задачи несбалансированного ОТ барицентра для непрерывных распределений общего вида в настоящее время отсутствуют. Существующие методы решения задачи несбалансированного ОТ барицентра, а также методы частичного ОТ, такие как SPOT и m-POT, либо ориентированы на работу с дискретными распределениями и, следовательно, обладают описанными выше ограничениями, либо применимы лишь к специальным классам непрерывных распределений.

Таким образом, актуальной задачей является разработка новых параметрических методов решения задачи несбалансированного ОТ, не требующих неустойчивой минимаксной оптимизации и обладающих строгими теоретическими гарантиями эффективности. Кроме того, необходима разработка масштабируемых численных методов, которые позволят находить решения задач несбалансированного ОТ барицентра и частичного ОТ.

Одной из ключевых математических задач является аналитический вывод двойственной формулировки для несбалансированного ОТ, позволяющей вычислять транспортные отображения с использованием легкой параметризации и без необходимости решения минимаксных задач. Другие связанные задачи включают аналитический вывод двойственных формулировок для задач несбалансированного ОТ барицентра и частичного ОТ, позволяющих вычислять соответствующие отображения с использованием выразительных параметризаций, например, нейросетей. Важными подзадачами являются получение теоретических результатов, устанавливающих связь между точностью нахождения оптимума двойственных функционалов и ошибкой аппроксимации решений прямых транспортных задач. Разработка теоретически обоснованных параметрических методов вычисления обобщенного ОТ позволит повысить эффективность их применения к описанным ранее задачам реального мира.

Целью диссертации является разработка математических моделей, основанных на нейросетях или легкой параметризации, для решения задач непрерывного обобщенного ОТ: вычисления несбалансированных транспортных отображений между непрерывными вероятностными распределениями, нахождения несбалансированных ОТ барицентров семейства непрерывных распределений и вычисления частичных транспортных отображений между вероятностными распределениями. Для достижения цели в работе были определены и решены следующие **задачи**:

1. Разработать численный метод, основанный на двойственной формулировке задачи несбалансированного оптимального

транспорта, для вычисления оптимального отображения между непрерывными распределениями с параметризацией посредством смесей гауссовых распределений; вывести теоретические оценки, устанавливающие связь между ошибками аппроксимации решений предложенных двойственной и прямой задач; получить теоретическую верхнюю оценку статистической ошибки для предложенной двойственной задачи.

2. Разработать численный метод, основанный на двойственной формулировке задачи несбалансированного оптимального транспорта, для вычисления (полу)несбалансированных ОТ барицентров семейства непрерывных вероятностных распределений с использованием нейронных сетей; вывести теоретические оценки, устанавливающие связь между ошибками аппроксимации решений предложенных двойственной и прямой задач.
3. Разработать численный метод, основанный на двойственной формулировке (полу)частичной задачи оптимального транспорта, для вычисления оптимального отображения между непрерывными распределениями с использованием нейронных сетей.

Научная новизна:

1. Впервые предложен теоретически обоснованный метод вычисления несбалансированных транспортных отображений между непрерывными распределениями, который является одноэтапным, не требует сложной минимаксной оптимизации и использует легкую параметризацию. Метод основан на новой идее применения специальной замены переменных в двойственной формулировке задачи несбалансированного оптимального транспорта и параметризации с помощью смесей гауссовых распределений. В отличие от существующих неминимаксных методов, основанных на решениях дискретной задачи несбалансированного ОТ, разработанный подход имеет строгие теоретические гарантии.
2. Впервые предложен масштабируемый и теоретически обоснованный метод вычисления (полу)несбалансированного ОТ барицентра для семейства непрерывных распределений. Метод основан на введении нового регуляризатора, обеспечивающего m-конгруэнтность решений двойственной задачи, и использовании нейронных сетей. Данный подход является первым методом, позволяющим вычислять несбалансированные ОТ барицентры для непрерывных распределений. Ранее эффективные численные методы вычисления несбалансированных ОТ барицентров в пространствах высокой размерности не существовали, все известные подходы рассматривали только дискретный случай.
3. Впервые предложен масштабируемый и теоретически обоснованный метод вычисления (полу)частичных транспортных отображений

между непрерывными распределениями. Данный метод является первым подходом, позволяющим вычислять (полу)частичные транспортные отображения для непрерывных распределений. Ранее эффективные численные методы решения (полу)частичных задач оптимального транспорта в высокоразмерных пространствах отсутствовали, все существующие подходы рассматривали только дискретный случай.

Теоретическая значимость. Разработанные подходы, основанные на численных методах обобщенного (несбалансированного и частичного) ОТ, позволяют эффективно решать ряд задач математического моделирования, для которых ранее не существовало альтернативных решений или существующие методы имели критические ограничения. А именно:

- Предложенный метод вычисления несбалансированных транспортных отображений позволяет устранять ограничения предыдущих подходов, такие как необходимость минимаксной оптимизации или отсутствие строгих теоретических обоснований;
- Предложенные подходы к вычислению несбалансированных барицентров и частичных транспортных отображений с использованием нейронных сетей являются первыми методами данного типа, применимыми к случаю непрерывных распределений.

Предложенные подходы позволяют проводить строгий теоретический анализ получаемых решений. Например, они дают возможность выводить оценки для статистических ошибок, характеризующих влияние конечного размера выборок на качество решений.

Практическая значимость работы заключается в том, что предложенные подходы математического моделирования, основанные на численных методах обобщенного ОТ, обладают широким спектром потенциальных приложений в прикладных областях. Например, разработанные методы несбалансированного и частичного ОТ протестированы в задачах преобразования синтетических и модельных данных и могут быть использованы в таких задачах компьютерного зрения, как обучение моделей для сверхразрешения и переноса стиля изображений. Другие области применения данных подходов связаны с задачами в области одноклеточной биологии, например, сопоставлением несогласованных распределений клеток на различных стадиях лечения онкологических заболеваний. Кроме того, разработанный метод вычисления несбалансированного ОТ барицентра имеет потенциал применения в задачах агрегации данных. Возможные приложения включают объединение данных, поступающих с МРТ-сканеров различных клиник, или данных геологических симуляторов, моделирующих различные аспекты строения подземных пластов. Такая агрегация может повысить эффективность моделей сегментации или классификации, построенных на этих данных.

Методология и методы исследования. Для решения задач, поставленных в работе, применяются методы теории несбалансированного

и частичного ОТ, позволяющие аналитически выводить двойственные формулировки обобщенных задач ОТ и устанавливать связи между решениями прямых и двойственных задач. Теоретическое обоснование разработанных подходов опирается на методы теории вероятностей, теории меры, статистики, линейной алгебры, функционального и выпуклого анализа, которые позволяют исследовать свойства предложенных оптимизационных функционалов, связи между решениями прямых и двойственных задач или выводить оценки для статистических ошибок.

Существенную роль в исследовании играют подходы машинного и глубинного обучения, используемые для параметризации функций, аппроксимирующих минимизаторы предложенных функционалов, посредством смесей гауссовских распределений или нейросетей. Для построения практических алгоритмов решения задач обобщенного ОТ применяются методы стохастической оптимизации. Практическая реализация и тестирование предложенных методов осуществлялись с использованием языка программирования Python и фреймворка PyTorch глубинного обучения. Оценка качества экспериментальных результатов выполнялась количественными и качественными методами, стандартными для области генеративного машинного обучения.

Основные положения, выносимые на защиту:

1. Разработан неминимаксный алгоритм для вычисления несбалансированных оптимальных транспортных отображений между непрерывными распределениями, основанный на использовании смесей гауссовых распределений для параметризации потенциалов в двойственной задаче. Выведены теоретические оценки, устанавливающие связь между ошибкой минимизации разработанного функционала оптимизации и ошибкой решения исходной задачи несбалансированного оптимального транспорта. Выведена теоретическая верхняя граница статистической ошибки для разработанного функционала оптимизации.
2. Разработан алгоритм вычисления (полу)несбалансированных ОТ барицентров для семейства непрерывных вероятностных распределений, основанный на использовании нейронных сетей и условия m-конгруэнтности двойственных переменных задачи несбалансированного барицентра. Выведены теоретические оценки, устанавливающие связь между ошибкой минимизации разработанного функционала оптимизации и ошибкой решения исходной задачи несбалансированного ОТ барицентра.
3. Разработан алгоритм вычисления (полу)частичных транспортных отображений между непрерывными распределениями, основанный на использовании нейронных сетей.

Полученные результаты соответствуют следующим пунктам паспорта специальности 1.2.2 «Математическое моделирование,

численные методы и комплексы программ»: **п.1** «Разработка новых математических методов моделирования объектов и явлений», **п.2** «Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий» и **п.3** «Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента.»

Достоверность полученных результатов обеспечивается корректным применением теории обобщенного ОТ и связанных с ней апробированных математических инструментов, а также строгим доказательством всех теоретических результатов. Предложенные алгоритмы протестированы на множестве синтетических и реальных данных. Подробные протоколы и открытый исходный код обеспечивают воспроизводимость экспериментов.

Разработанные численные методы и алгоритмы реализованы в виде комплекса проблемно-ориентированных программ для решения задач обобщенного ОТ. Комплекс обеспечивает обучение параметрических моделей, построение несбалансированных и (полу)частичных ОТ отображений и (полу)несбалансированных ОТ барицентров, а также оценку полученных решений. Программный код исследования публично доступен в репозитории автора на GitHub: `https://github.com/milenagazdieva`.

Все результаты, выносимые на защиту, опубликованы в рецензируемых научных журналах и трудах ведущих международных конференций.

Публикации. Список публикаций автора приведен в конце диссертации. Основные результаты по теме диссертации изложены в **5** работах [1-5], включая **2** работы в изданиях из списка ВАК и приравненных к ним, из которых **1** опубликована в международном журнале уровня **SJR rank Q1** и индексируется в базе данных **Scopus**. Другие **3** работы опубликованы в трудах ведущих международных конференций уровня **Core rank A*** и также индексируются в **Scopus**.

Апробация. Результаты диссертационной работы были представлены на **3** ведущих международных научных конференциях уровня **Core A*** по машинному обучению и искусственному интеллекту:

1. The Thirty-seventh Annual Conference on Neural Information Processing Systems (NeurIPS), 2023, Новый Орлеан, США;
2. The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024, Ванкувер, Канада;
3. The Thirteenth International Conference on Learning Representations (ICLR), 2025, Сингапур.

Кроме того, результаты диссертации были представлены на следующих **8** научных семинарах, лекциях и локальных конференциях:

- Семинар института AIRI, 2023, Москва, Россия;
- Конференция Fall into ML в НИУ ВШЭ, 2023, Москва, Россия;
- Семинар института AIRI, 2024, Москва, Россия;

– Конференция Fall into ML в НИУ ВШЭ, 2024, Москва, Россия;
– Конференция DataFest в Яндексе, 2025, Москва, Россия;
– Конференция Fall into ML в НИУ ВШЭ, 2025, Москва, Россия;
– Лекция в Beijing Institute of Mathematical Sciences and Applications, 2025, Пекин, Китай;
– Научный семинар лаборатории LARSS Сколковского Института Науки и Технологий, 2025, Москва, Россия.

Часть результатов диссертационной работы, а именно разработанный алгоритм вычисления частичных транспортных отображений, прошла апробацию и была внедрена в платформу для мониторинга и прогнозирования природных катастроф на территории Российской Федерации. Данная платформа разработана Центром искусственного интеллекта Сколковского института науки и технологий, в котором работает автор диссертации. Кроме того, за преподавание курсов по генеративным моделям, основанных в том числе на научных результатах данной диссертации, соискатель в 2025 году была удостоена премии Yandex ML Prize от компании Яндекс в номинации "Начинающие преподаватели".

Интерес автора диссертации к рассматриваемой тематике начался с работы [1], в которой исследовалось применение различных подходов глубинного обучения к задаче сверхразрешения (SR) изображений в условиях несогласованных обучающих данных. В ходе исследования было показано, что многие нейросетевые алгоритмы для сверхразрешения могут быть интерпретированы как методы решения задачи несбалансированного ОТ. Дальнейшие исследования были посвящены углубленному изучению несбалансированного и частичного (т.е. обобщенного) ОТ, а ключевые результаты, составляющие основу данной диссертации, были получены в работах [3-5]. Кроме того, в статье [2] был проведен сравнительный анализ методов классического ОТ и разработанных методов обобщенного ОТ.

Личный вклад. Все результаты, выносимые на защиту, получены лично автором. Все алгоритмы и теоретические результаты, опубликованные в работах [3-5] и представленные в диссертации, разработаны и доказаны автором диссертации. Диссертация включает результаты совместной работы: экспериментальное тестирование метода вычисления полунесбалансированного ОТ барицентра [4] выполнено в сотрудничестве с соавтором Дж. Чои; автор диссертации разработала подход для вычисления получастичных транспортных отображений [5], осуществила его предварительную реализацию и протестировала при различных параметрах, тогда как Д. Селиханович адаптировал подход для сравнения с альтернативными методами. Все ключевые эксперименты в работах [1,3] выполнены автором диссертации; вспомогательные эксперименты в этих работах проведены соавторами А. Асадулаевым, Л. Роут и А. Коротиным. Теоретические результаты, представленные в статье [1], получены в соавторстве с П. Мокровым и А. Коротиным. Анализ методов в статье [2] проведен лично автором.

Замечание. Диссертационная работа выполнена автором в автономной некоммерческой образовательной организации высшего образования «Сколковский институт науки и технологий».

Объем и структура работы. Диссертация состоит из введения, 4 глав и заключения. Полный объем работы составляет 176 страниц, включая 38 рисунков и 39 таблиц. Библиография состоит из 152 наименований.

Содержание работы

Во **введении** обосновывается актуальность проведенного исследования, формулируются цель и задачи работы, излагается научная новизна, а также теоретическая и практическая значимости представленных результатов.

Первая глава содержит общий обзор основных задач, решаемых в диссертации: вычисление несбалансированных OT отображений (**задача 1**), вычисление (полу)несбалансированных OT барицентров (**задача 2**) и вычисление (полу)частичных OT отображений (**задача 3**). Ниже приведены математические формулировки задач, а также краткое обсуждение ограничений существующих подходов.

Для общности изложения задачи диссертации формулируются в предположении, что рассматриваемые пространства являются польскими пространствами. На практике, как правило, используются евклидовы пространства или их подмножества; детали предположений по каждой из задач приводятся в последующих главах работы.

Далее D-мерное евклидово пространство обозначается как $(\mathbb{R}^D, \|\cdot\|)$, где $\|\cdot\|$ обозначает евклидову норму; общие польские пространства обозначаются как $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$. Множество $\mathcal{P}_2(\mathcal{X})$ обозначает множество вероятностных мер на $\mathcal{X}$ с конечными вторыми моментами. Кроме того, множество $\mathcal{M}_{2,+}(\mathcal{X})$ обозначает неотрицательные меры на $\mathcal{X}$ с конечными вторыми моментами. Пространство всех *непрерывных* функций на $\mathcal{X}$ обозначается как $\mathcal{C}(\mathcal{X})$. Для заданной меры $\gamma \in \mathcal{M}_{2,+}(\mathcal{X}\times\mathcal{Y})$ ее полная масса обозначается как $\|\gamma\|_1 \stackrel{\text{def}}{=} \int_{\mathcal{X}\times\mathcal{Y}} d\gamma(x,y)$, ее маргиналы обозначаются как γ_x, γ_y. Кроме того, $\gamma(\cdot|x)$ обозначает условную *вероятностную* меру. Мы используем $\overline{f}$ для обозначения сопряженной по Фенхелю функции f: $\overline{f}(t) \stackrel{\text{def}}{=} \sup_{u\in\mathbb{R}}\{ut - f(u)\}$; $\mathbb{I}_A$ – для обозначения выпуклого индикатора множества A, т.е. $\mathbb{I}_A(x) = 0$, если $x \in A$; $\mathbb{I}_A(x) = +\infty$, если $x \notin A$. Образ распределения $\mathbb{P} \in \mathcal{P}_2(\mathbb{R}^D)$ при действии измеримого отображения $T : \mathbb{R}^D \to \mathbb{R}^D$ обозначается как $T\sharp\mathbb{P}$.

Для *положительных* мер μ_1,$\mu_2 \in \mathcal{M}_{2,+}(\mathcal{X})$ и полунепрерывной снизу функции $f : \mathbb{R} \to \mathbb{R}\cup\{\infty\}$ **f-дивергенция** между μ_1,μ_2 определяется как

$$D_f(\mu_1\|\mu_2) \stackrel{\text{def}}{=} \int_{\mathcal{X}} f\left(\frac{d\mu_1}{d\mu_2}(x)\right) d\mu_2(x), \text{ если } \mu_1 \ll \mu_2, \text{ и } +\infty \text{ иначе.} \quad (1)$$

Для *положительной меры* $\mu \in \mathcal{M}_{2,+}(\mathbb{R}^D)$ ее энтропия задается как

$$H(\mu) \stackrel{\text{def}}{=} -\int_{\mathbb{R}^D} \mu(x)\log\mu(x)\,dx + \|\mu\|_1, \text{ если } \mu \text{ - a.c., и } -\infty \text{ иначе.} \quad (2)$$

Задача 1 (Вычисление несбалансированного оптимального транспорта)**.** *Для пары вероятностных распределений* $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ *задача несбалансированного OT (unbalanced optimal transport, UOT) задается как*

$$UOT_{c,\psi,\phi}(\mathbb{P},\mathbb{Q}) \stackrel{def}{=} \inf_{\gamma \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma(x,y) + D_\psi\left(\gamma_x \| \mathbb{P}\right) + D_\phi\left(\gamma_y \| \mathbb{Q}\right), \quad (3)$$

где $c(x,y) \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})$ *— непрерывная функция стоимости, а* D_ψ *и* D_ϕ *— это* f*-дивергенции на пространствах* $\mathcal{X}$ *и* $\mathcal{Y}$ *соответственно, см. рис. 1а.*

Решение задачи (3) *состоит в нахождении меры* $\gamma^* \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})$*, доставляющей минимум. Такая мера называется UOT планом, а соответствующее минимальное значение* (3) *— стоимостью UOT. При этом UOT план* γ^* *определяет стохастическое или детерминистическое отображение* $T^* : x \mapsto \gamma^*(\cdot \mid x)$*, называемое UOT отображением.*

Заметим, что при $\phi(x) = \psi(x) = \mathbb{I}_{x=1}$ маргиналы плана должны удовлетворять условиям $\gamma_x = \mathbb{P}, \gamma_y = \mathbb{Q}$ и задача UOT (3) сводится к классической задаче OT. Таким образом, UOT ослабляет ограничения на маргиналы плана и позволяет повысить устойчивость решений к неточностям в данных. Если $\phi(x) = \mathbb{I}_{x=1}$, то фиксируется второй маргинал плана $\gamma_y = \mathbb{Q}$ и получается задача полунесбалансированного OT (semi-unbalanced OT, SUOT). Задача несбалансированного *энтропийного* OT (UEOT) может быть получена добавлением энтропийного регуляризатора $-\varepsilon H(\gamma)$ к функционалу (3). Отметим, что хотя ниже рассматриваются основные свойства задачи UOT, эти свойства выполняются и для задачи UEOT.

Формулировка (3) задачи несбалансированного оптимального транспорта представляет собой оптимизацию (возможно, нелинейного) функционала от γ в пространстве неотрицательных мер на $\mathcal{X}\times\mathcal{Y}$. Заметим, что для пар $\mathbb{P}$, $\mathbb{Q}$ общего вида аналитические решения (3) неизвестны, а решение данной оптимизационной задачи на практике вычислительно затруднено ввиду нетривиальных членов дивергенции.

Для **дискретных** распределений $\mathbb{P}$ и $\mathbb{Q}$ с конечным носителем задача обычно решается методами линейного программирования или их регуляризованными вариантами, что приводит к итерационным матричным алгоритмам. В то же время, при работе с **непрерывными** распределениями, доступными лишь через выборки, дискретные методы теряют эффективность, поскольку построение требуемых точных дискретизаций распределений становится вычислительно неосуществимым.

Далее мы приведем краткий обзор основных существующих методов решения задачи несбалансированного OT (3) для непрерывных распределений $\mathbb{P}$ и $\mathbb{Q}$ и выделим их ключевые ограничения, устраняемые в настоящей работе. Многие современные методы, направленные на решение задачи UOT для непрерывных распределений, в частности NubOT, UOT-FM и GENOT, используют решения дискретной задачи UOT при построении своих алгоритмов, при этом теоретическая обоснованность решений таких методов остается недостаточно изученной.

Существующие непрерывные методы UOT основаны на решении двойственной задачи к (3) и последующем восстановлении UOT отображения T^* (детерминированного или стохастического) по оцененным двойственным переменным. Для $\mathbb{P} \in \mathcal{P}_2(\mathcal{X})$, $\mathbb{Q} \in \mathcal{P}_2(\mathcal{Y})$ **двойственная формулировка** задачи UOT имеет следующий вид:

$$\mathrm{UOT}_{c,\psi,\phi}(\mathbb{P},\mathbb{Q}) = \sup_{g \oplus f \leq c(\cdot,\cdot)} \left\{ -\int_{\mathcal{X}} \overline{\psi}(-g(x))d\mathbb{P}(x) - \int_{\mathcal{Y}} \overline{\phi}(-f(y))d\mathbb{Q}(y) \right\}, \qquad (4)$$

где супремум берется по $\mathbb{P}$-интегрируемым функциям $g : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ и $\mathbb{Q}$-интегрируемым $f : \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ таким, что $\forall x \in \mathcal{X}, y \in \mathcal{Y} :\ g(x) + f(y) \leq c(x,y)$. Эти функции обычно называются *потенциалами*.

Кроме того, существует альтернативная к (4) *полудвойственная формулировка* UOT задачи:

$$\mathrm{UOT}_{c,\psi,\phi}(\mathbb{P},\mathbb{Q}) = \sup_{f \in \mathcal{C}(\mathcal{Y})} \left\{ -\int_{\mathcal{X}} \overline{\psi}(-f^c(x))d\mathbb{P}(x) - \int_{\mathcal{Y}} \overline{\phi}(-f(y))d\mathbb{Q}(y) \right\}, \qquad (5)$$

где супремум берется по непрерывным функциям (потенциалам) f, а $f^c(x)$ обозначает *c-преобразование*: $f^c(x) \stackrel{\text{def}}{=} \inf_{y \in \mathcal{Y}} \big(c(x,y) - f(y)\big)$. При этом для оптимального потенциала f^* и UOT-плана γ^* выполняется $\operatorname{supp} \gamma^*(\cdot \mid x) \subseteq \arg\min_{y \in \mathcal{Y}} \big(c(x,y) - f^*(y)\big)$ для γ^*_x-почти всех x. При стандартных предположениях, обеспечивающих единственность и детерминированность γ^*, это позволяет восстановить UOT отображение $T^*(x) = \arg\min_{y \in \mathcal{Y}} \big(c(x,y) - f^*(y)\big)$, а также маргинал γ^*_x, по оптимальному потенциалу f^*. Этот факт позволяет решать задачу (3), сосредоточившись только на решении двойственной или полудвойственной задач (4), (5), которые представляются более простыми для решения, поскольку устраняют необходимость в нетривиальной минимизации дивергенций в прямой задаче UOT (3). Однако решение задачи (5), то есть нахождение двойственных потенциалов, остается сложным, поскольку вычисление c-преобразования f^c само по себе представляет задачу оптимизации.

В методе UOTM данная оптимизационная задача поиска седловой точки решается с использованием двух нейросетей, параметризующих потенциал f и транспортное отображение T и обучаемых с помощью алгоритма градиентного спуска-подъема. В ином подходе к решению задачи UOT вводится дополнительная параметризация для коэффициента масштабирования меры $\mathbb{P}$, то есть всего рассматриваются три нейросети. В результате данные методы опираются на сложную нейросетевую параметризацию, могут страдать от численной нестабильности в процессе обучения и требуют тщательного подбора гиперпараметров.

В данной диссертации **предлагается** новый **численный метод** (глава 2 работы), который преодолевает ограничения существующих подходов, а именно, имеет строгие теоретические обоснования, основан на легкой параметризации с помощью смеси гауссовских распределений и обладает неминимаксным целевым функционалом.

Задача 2 (Задача поиска полунесбалансированного OT барицентра)**.** *Рассмотрим вероятностные меры* $\mathbb{P}_k \in \mathcal{P}(\mathcal{X}_k)$ *и непрерывные функции стоимости* $c_k(x,y): \mathcal{X}_k \times \mathcal{Y} \mapsto \mathbb{R}$, $k \in \overline{K}$. *Пусть заданы веса* $\lambda_k \geq 0$ *такие, что* $\sum_{k=1}^{K} \lambda_k = 1$, *и набор* f*-дивергенций* $\psi_{[1:K]}$. *Задача полунесбалансированного OT барицентра состоит в нахождении минимизатора взвешенной суммы задач полунесбалансированного OT (SUOT), см. рис. 1б:*

$$\mathcal{B}_u^* \stackrel{def}{=} \inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{Y})} \mathcal{B}_u(\mathbb{Q}) \stackrel{def}{=} \inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{Y})} \sum_{k=1}^{K} \lambda_k SUOT_{c_k, \psi_k}(\mathbb{P}_k, \mathbb{Q}). \tag{6}$$

Мера $\mathbb{Q}^*$, *доставляющая минимум в задаче* (6) *(если она существует), называется SUOT барицентром,* $\mathcal{B}_u$ *называется функционалом SUOT барицентра, а его минимальное значение обозначается как* $\mathcal{B}_u^*$.

В настоящее время не существует эффективных подходов для вычисления полунесбалансированных OT барицентров для семейства непрерывных распределений и f-дивергенций общего вида. Существующие методы являются либо дискретными, либо решают классическую задачу OT барицентра, которая представляет собой частный случай задачи SUOT барицентра при $\psi_k \equiv \mathbb{I}_{x=1}$ $(\forall k \in \overline{K})$. Дискретные подходы страдают от проблем, которые обсуждались ранее в данном разделе. А именно, в случае, когда вероятностные распределения $\mathbb{P}_1, ..., \mathbb{P}_k$ являются высокоразмерными непрерывными вероятностными распределениями, дискретные методы вычисления SUOT барицентра не обеспечивают достаточного качества, поскольку требуемые точные дискретизации распределений вычислительно недостижимы. В то же время непрерывные методы решения задачи OT барицентра, такие как WIN, методы на основе ICNN и NOT, а также Energy-Guided подход, чувствительны к наличию нежелательного шума, выбросов и прочих неточностей во входных выборках данных.

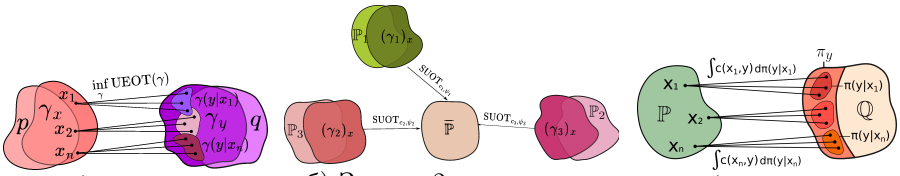
В данной диссертации **предлагается** первый **численный метод** (глава 3 данной работы) для вычисления полунесбалансированного OT барицентра для семейства непрерывных распределений.

Задача 3 (Вычисление неполного оптимального транспорта)**.** *Рассмотрим конечный параметр* $w \geq 1$. *Для заданных вероятностных мер* $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ *и* $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ *задача неполного OT (incomplete transport, IT) определяется как*

$$Cost_w(\mathbb{P}, \mathbb{Q}) \stackrel{def}{=} \inf_{\pi \in \Pi^w(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{Y}} c(x,y) d\pi(x,y), \tag{7}$$

где инфимум берется по множеству $\Pi^w(\mathbb{P}, \mathbb{Q})$ *вероятностных мер* $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, *первый маргинал которых удовлетворяет равенству* $\pi_x = \mathbb{P}$, *а второй — неравенству* $\pi_y \leq w\mathbb{Q}$ *(см. рис. 1в).*

В диссертации объектом исследования является задача получастичного OT с ослабленным вторым маргиналом, обозначаемая как задача *неполного* OT (7). Такая задача имеет дополнительную практическую



а) Задача 1: вычисление UOT плана.

б) Задача 2: вычисление SUOT барицентра $\overline{P}$ семейства $\{\mathbb{P}_k\}$.

в) Задача 3: вычисление неполного транспортного плана.

Рис. 1 — Схематическая иллюстрация задач 1, 2 и 3.

ценность, поскольку предоставляет способ контроля и повышения степени сходства между сгенерированными объектами и объектами из исходного распределения $\mathbb{P}$ путем варьирования параметра ω.

В частном случае $w = 1$ задача (7) сводится к классической задаче OT, так как неравенства могут быть заменены равенствами. Однако при других значениях $\omega > 1$ формулировка получастичного OT позволяет вычислять OT отображения между исходным распределением $\mathbb{P}$ и определенной частью $(1/\omega)$ целевого распределения $\mathbb{Q}$, контролируемой параметром ω. С другой стороны, получастичный OT представляет собой частный случай формулировки частичного OT, которая формализует задачу поиска OT планов между частями исходного и целевого распределений:

$$\inf_{\substack{m\pi_x \le w_0\mathbb{P} \\ m\pi_y \le w_1\mathbb{Q}}} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\big[m\pi\big](x,y). \tag{8}$$

Здесь $w_0, w_1 \ge m \ge 0$ — конечные параметры, а инфимум берется по мерам $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$, удовлетворяющим неравенствам $m\pi_x \le w_0\mathbb{P}$ и $m\pi_y \le w_1\mathbb{Q}$.

Научный и практический интерес к вычислению частичных транспортных планов состоит в том, что задача частичного OT представляет собой еще один способ обобщения классической формулировки OT и также позволяет работать с данными, содержащими шум, выбросы или прочие неточности. Важное отличие формулировки частичного OT от несбалансированного OT состоит в том, что в частичном OT явно фиксируются распределения, между которыми вычисляется транспортный план, тогда как в UOT эти распределения не определены однозначно.

Следует отметить, что нахождение неполного OT отображения (условного плана) представляет собой сложную задачу, поскольку требует разработки метода, обеспечивающего выполнение условия неравенства для сгенерированного и целевого распределений. В настоящее время существуют только дискретные подходы для вычисления частичных OT отображений между распределениями $\mathbb{P}$ и $\mathbb{Q}$, которые, как отмечалось ранее, не обеспечивают достаточного качества в случае непрерывных распределений $\mathbb{P}$ и $\mathbb{Q}$.

В данной диссертации предлагается первый **численный метод** (глава 4 настоящей работы) для вычисления неполных (получастичных) OT отображений между непрерывными распределениями.

Основные результаты глав 2–4 приведены ниже. Каждая глава посвящена решению одной из задач обобщенного OT 1, 2, 3.

Вторая глава посвящена задаче нахождения несбалансированного оптимального транспортного отображения между двумя непрерывными вероятностными распределениями (задача 1). Для решения этой задачи предлагается и теоретически обосновывается численный метод, основанный на параметризации с помощью смеси гауссовских распределений. Результаты данной главы подробно изложены ниже.

В данной главе мы работаем с евклидовыми пространствами $\mathcal{X}=\mathcal{Y}=(\mathbb{R}^D,\|\cdot\|)$ и рассматриваем аналог задачи (3) с энтропийной регуляризацией и квадратичной функцией стоимости $c(x,y)=\frac{\|x-y\|^2}{2}$. Кроме того, мы рассматриваем только **абсолютно непрерывные** вероятностные меры с конечной дифференциальной энтропией, которые обозначаются как $p,q\in\mathcal{P}_{2,ac}(\mathbb{R}^d)$. Для фиксированного $\varepsilon>0$ мы рассматриваем задачу несбалансированного EOT (unbalanced EOT, UEOT) между p и q, которая состоит в нахождении минимизатора функционала:

$$\inf_{\gamma\in\mathcal{M}_{2,+}(\mathbb{R}^d\times\mathbb{R}^d)}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\frac{\|x-y\|^2}{2}\gamma(x,y)dxdy-\varepsilon H(\gamma)+D_{f_1}\left(\gamma_x\|p\right)+D_{f_2}\left(\gamma_y\|q\right). \quad (9)$$

Инфимум в задаче (9) достигается для абсолютно непрерывной меры γ^* благодаря наличию слагаемого $-\varepsilon H(\gamma)$. Для вычисления несбалансированного EOT плана γ^* в (9) предлагается параметризовать

$$\gamma_{\theta,w}(x,y)\stackrel{\text{def}}{=}u_\omega(x)\frac{\exp\{\langle x,y\rangle/\varepsilon\}v_\theta(y)}{c_\theta(x)}, \quad (10)$$

где функции $u_\omega{:}\mathbb{R}^D\to\mathbb{R}$, $v_\theta{:}\mathbb{R}^D\to\mathbb{R}$ имеют независимые наборы параметров ω,θ соответственно, а $c_\theta(x)\stackrel{\text{def}}{=}\int_{\mathbb{R}^d}\exp\{\frac{\langle x,y\rangle}{\varepsilon}\}v_\theta(y)dy$ является нормировочной константой. Затем для нахождения приближенного плана $\gamma_{\theta,\omega}$ предлагается рассмотреть следующую неминимаксную задачу оптимизации:

$$\min_{\theta,\,\omega}\mathcal{L}(\theta,\omega)\stackrel{\text{def}}{=}\min_{\theta,\,\omega}\left\{\int_{\mathbb{R}^d}\overline{f}_1(-\varepsilon\log\frac{u_\omega(x)}{c_\theta(x)}-\frac{\|x\|^2}{2})p(x)dx+\right.$$
$$\left.\int_{\mathbb{R}^d}\overline{f}_2(-\varepsilon\log v_\theta(y)-\frac{\|y\|^2}{2})q(y)dy+\varepsilon\|u_\omega\|_1\right\}. \quad (11)$$

Ключевая особенность предложенной задачи заключается в том, что она позволяет минимизировать дивергенцию $\mathrm{D_{KL}}$ между приближенным планом $\gamma_{\theta,\omega}$ и истинным планом γ^* без необходимости прямого доступа к последнему. В диссертации доказывается следующая теорема, демонстрирующая связь между ошибкой решения задачи (11) и ошибкой аппроксимации решения прямой задачи UEOT (9) посредством параметризованного плана $\gamma_{\theta,\omega}$.

Теорема 1 (Вычислительно удобная формулировка минимизации $\mathrm{D_{KL}}$)**.** *Предположим, что план γ^* параметризован с использованием* (10)*. Тогда выполняется следующее неравенство: $\varepsilon D_{KL}\left(\gamma^*\|\gamma_{\theta,\omega}\right)\leq\mathcal{L}(\theta,\omega)-\mathcal{L}^*$, где $\mathcal{L}(\theta,\omega)$ задается выражением* (11)*, а константа $(-\mathcal{L}^*)$ является оптимальным значением двойственной задачи UEOT:*

$$\sup_{(\phi,\psi)}\left\{-\varepsilon\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\exp\{\frac{1}{\varepsilon}(\phi(x)+\psi(y)-\frac{\|x-y\|^2}{2})\}dxdy-\right.$$

$$\left.\int_{\mathbb{R}^d}\overline{f}_1(-\phi(x))p(x)dx-\int_{\mathbb{R}^d}\overline{f}_2(-\psi(y))q(y)dy\right\},\tag{12}$$

где супремум берется по измеримым функциям ϕ, ψ (потенциалам). Это неравенство является ***точным*** *в том смысле, что оно обращается в $0=0$ при $v_\theta(y)=\exp\{2\psi^*(y)-\|y\|^2/2\varepsilon\}$ и $u_\omega(x)=c_\theta(x)\exp\{2\phi^*(x)-\|x\|^2/2\varepsilon\}$.*

Данный результат можно интерпретировать как верхнюю оценку качества приближенного решения $\gamma_{\theta,\omega}$ задачи (9), восстановленного из приближенного решения двойственной задачи (12). Таким образом, теорема показывает, что вместо прямой минимизации $\mathrm{D_{KL}}$ для получения $\gamma_{\theta,\omega}$ можно оптимизировать верхнюю оценку (11), которая является вычислительно удобной. Действительно, функционал (11) представляет собой сумму математических ожиданий относительно вероятностных мер p, q, что позволяет использовать оценку Монте-Карло выражения (11) по случайным выборкам. Основная **сложность** здесь заключается в вычислении переменной c_θ и слагаемого $\|u_\omega\|_1$.

Для решения данной проблемы предлагается рассмотреть параметризацию в виде ненормированной смеси гауссовских распределений для $u_\omega(x)$ и $v_\theta(y)$:

$$v_\theta(y)=\sum_{k=1}^{K}\alpha_k\mathcal{N}(y|r_k,\varepsilon S_k);\quad u_\omega(x)=\sum_{l=1}^{L}\beta_l\mathcal{N}(x|\mu_l,\varepsilon\Sigma_l).\tag{13}$$

Здесь $\theta\stackrel{\text{def}}{=}\{\alpha_k,r_k,S_k\}_{k=1}^K$, $w\stackrel{\text{def}}{=}\{\beta_l,\mu_l,\Sigma_l\}_{l=1}^L$, где $\alpha_k,\beta_l\geq 0$, $r_k,\mu_l\in\mathbb{R}^d$ и $0\prec S_k,\Sigma_l\in\mathbb{R}^{d\times d}$. Ковариационные матрицы масштабируются на ε исключительно для удобства. Для данного типа параметризации существуют аналитические выражения как для $\|u_\omega\|_1$, так и для нормировочной константы $c_\theta(x)$.

В результате в работе предлагается **вычислительный алгоритм** для минимизации оценки Монте-Карло разработанного функционала (11), основанный на параметризации u_ω и v_θ с помощью смеси гауссовских распределений. Для оптимизации их параметров θ и ω предлагается использовать процедуру стохастического градиентного спуска с использованием случайных выборок из входных распределений p и q.

С теоретической точки зрения для восстановления UEOT плана γ^* необходимо решить задачу $\mathcal{L}(\theta,\omega)\to\min_{\theta,\omega}$, как указано в теореме 1. Однако на практике меры p и q доступны лишь через эмпирические выборки

$X \stackrel{\text{def}}{=} \{x_1,...,x_N\}$ и $Y \stackrel{\text{def}}{=} \{y_1,...,y_M\}$, поэтому требуется оптимизировать эмпирический аналог функционала $\mathcal{L}(\theta,\omega)$:

$$\widehat{\mathcal{L}}(\theta,w) \stackrel{\text{def}}{=} \left\{ \frac{1}{N}\sum_{i=1}^{N} \overline{f}_1(-\varepsilon \log \frac{u_\omega(x_i)}{c_\theta(x_i)} - \frac{\|x_i\|^2}{2}) + \right.$$
$$\left. \frac{1}{M}\sum_{j=1}^{M} \overline{f}_2(-\varepsilon \log v_\theta(y_j) - \frac{\|y_j\|^2}{2}) + \varepsilon \|u_\omega\|_1 \right\}. \quad (14)$$

Кроме того, доступные потенциалы u_ω и v_θ, по которым производится оптимизация функционала, принадлежат ограниченному классу функций. В частности, u_ω и v_θ параметризуются как ненормированные смеси гауссовских распределений с L и K компонентами соответственно. Это естественным образом порождает задачу количественной оценки того, *насколько точно восстановленный план* $\gamma_{\widehat{\theta},\widehat{\omega}}$ *аппроксимирует истинный UEOT план* γ^* *при условии, что* $(\widehat{\theta},\widehat{\omega}) = \arg\min_{\theta,\omega} \widehat{\mathcal{L}}(\theta,\omega)$. Для решения данной задачи в работе исследуется *ошибка обобщения* $\mathbb{E}\mathrm{D}_{\mathrm{KL}}\left(\gamma^* \| \gamma_{\widehat{\theta},\widehat{\omega}}\right)$, то есть математическое ожидание D_{KL} между γ^* и $\gamma_{\widehat{\theta},\widehat{\omega}}$, взятое относительно случайной реализации обучающих выборок $X \sim p$, $Y \sim q$.

Из теоремы 1 следует, что величина $\mathbb{E}\mathrm{D}_{\mathrm{KL}}\left(\gamma^* \| \gamma_{\widehat{\theta},\widehat{\omega}}\right)$ может быть ограничена сверху через $\mathbb{E}(\mathcal{L}(\widehat{\theta},\widehat{\omega}) - \mathcal{L}^*)$. При этом последняя величина может быть разложена на статистическую ошибку и ошибку аппроксимации:

$$\mathbb{E}(\mathcal{L}(\widehat{\theta},\widehat{\omega}) - \mathcal{L}^*) = \mathbb{E}[\mathcal{L}(\widehat{\theta},\widehat{\omega}) - \mathcal{L}(\overline{\theta},\overline{\omega})] + \mathbb{E}[\mathcal{L}(\overline{\theta},\overline{\omega}) - \mathcal{L}^*] =$$
$$\underbrace{\mathbb{E}[\mathcal{L}(\widehat{\theta},\widehat{\omega}) - \mathcal{L}(\overline{\theta},\overline{\omega})]}_{\text{Стат. ошибка}} + \underbrace{[\mathcal{L}(\overline{\theta},\overline{\omega}) - \mathcal{L}^*]}_{\text{Ошибка аппрокс.}}. \quad (15)$$

Здесь $(\overline{\theta},\overline{\omega}) = \arg\min_{(\theta,\omega)\in\Theta\times\Omega} \mathcal{L}(\theta,\omega)$ обозначают параметры, которые обеспечивают наилучшую аппроксимацию $\mathcal{L}(\theta,\omega)$ в допустимом классе. В диссертации доказывается следующее утверждение, устанавливающее количественную верхнюю оценку для статистической ошибки.

Предложение 2 (Оценка сверху для статистической ошибки)**.** *Пусть* p,q *имеют компактные носители и функции* $\overline{f}_1$, $\overline{f}_2$ *являются липшицевыми. Предположим, что рассматриваемые параметрические классы* Θ, Ω $(\ni (\theta,\omega))$ *состоят из ненормированных смесей гауссовских распределений с* K *и* L *компонентами соответственно, ограниченными средними* $\|r_k\|,\|\mu_l\| \le R$, *ковариациями* $sI \preceq S_k,\Sigma_l \preceq SI$ *и весами* $a \le \alpha_k,\beta_l \le A$ *(для некоторых* $R > 0$, $0 < s \le S$ *и* $0 < a \le A$*). Тогда выполняется следующее неравенство:*

$$\mathbb{E}\big[\mathcal{L}(\widehat{\theta},\widehat{\omega}) - \mathcal{L}(\overline{\theta},\overline{\omega})\big] \le O(\frac{1}{\sqrt{N}}) + O(\frac{1}{\sqrt{M}}),$$

где $O(\cdot)$ *скрывает константы, зависящие только от* $K,L,R,s,S,a,A,p,q,\varepsilon$, *но не от размеров выборок* M,N.

Данное утверждение показывает, что статистическая ошибка стремится к нулю при $N,M \to \infty$ со стандартной параметрической скоростью. Открытым остается следующий вопрос: *может ли ошибка аппроксимации быть сделана сколь угодно малой*? Приведенная ниже теорема дает положительный ответ на данный вопрос.

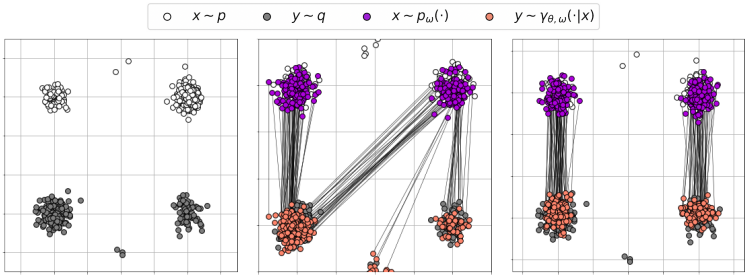
Теорема 3 (Параметризация переменных смесью гауссовских распределений обеспечивает универсальную аппроксимацию UEOT планов)**.** *Пусть p и q имеют компактные носители, а функции $\overline{f}_1$, $\overline{f}_2$ являются липшицевыми. Тогда для любого $\delta > 0$ существуют смеси гауссовских распределений v_θ, u_ω (13) со* ***скалярными*** *ковариациями компонент $S_k = \lambda_k I_d \succ 0$, $\Sigma_l = \zeta_l I_d \succ 0$, удовлетворяющие неравенству $D_{KL}\left(\gamma^* \| \gamma_{\theta,\omega}\right) \leq \varepsilon^{-1}(\mathcal{L}(\theta,\omega) - \mathcal{L}^*) < \delta$.*

В совокупности результаты предложения 2 и теоремы 3 показывают, что ошибку обобщения можно сделать *сколь угодно малой* при достаточно большом размере выборок и числе компонент в гауссовской параметризации. Это означает, что с теоретической точки зрения разработанный метод способен восстанавливать UEOT план с произвольно высокой точностью.

Для эмпирической проверки эффективности предложенного метода, оптимизирующего функционал (11) с использованием смесей гауссовских распределений, в работе проведена серия вычислительных экспериментов. Приведены как качественные, так и количественные оценки UEOT отображений, вычисленных с помощью разработанного алгоритма. С этой целью рассмотрен ряд задач и приложений на синтетических и реальных данных. В частности, рассмотрена задача вычисления UEOT отображений: (а) между двумерными смесями гауссовских распределений, страдающими от проблем дисбаланса классов и выбросов; (б) для переноса массы в латентном пространстве автокодировщиков данных.

Пример 1 (UEOT между смесями гауссовских распределений)**.** *Ниже приведены качественные экспериментальные результаты восстановления UEOT планов между случайно инициализированными смесями двумерных гауссовских распределений, страдающих от дисбаланса классов и выбросов в исходных и целевых мерах p, q. Разработанный метод (U-LightOT) реализован с использованием дивергенции D_{KL}. Результаты, см. рис. 2, показывают, что метод устойчив к указанным проблемам, тогда как альтернативный метод LightSB, восстанавливающий сбалансированные EOT планы между мерами, не способен с ними справиться.*

Третья глава посвящена задаче нахождения полунесбалансированного OT барицентра семейства непрерывных вероятностных распределений. Для решения этой задачи предлагается и теоретически обосновывается численный метод, основанный на использовании нейронных сетей и условия m-конгруэнтности в двойственной задаче.



а) Меры p,q б) LightSB в) U-LightOT (ours)

Рис. 2 — Условные планы $\gamma_{\theta,\omega}(y|x)$, обученные разработанным методом и методом LightSB в эксперименте *«Gaussian Mixtures with outliers»*.

В данной главе рассматриваются компактные подмножества евклидовых пространств: $\mathcal{X} \subset \mathbb{R}^{D'}$, $\mathcal{X}_k \subset \mathbb{R}^{D_k}$, $\mathcal{Y} \subset \mathbb{R}^D$. При этих предположениях SUOT барицентр $\mathbb{Q}^*$, доставляющий минимум в задаче (6), всегда существует, но его единственность не гарантирована.

Очевидный способ вывода двойственной задачи поиска SUOT барицентра состоит в подстановке двойственной (полу)несбалансированной OT задачи (4) в функционал прямой задачи SUOT барицентра (6), что приводит к сложной задаче типа min-max-min. Предлагаемый алгоритм основан на **max-min реформулировке** двойственной задачи поиска (полу)несбалансированного OT барицентра (6). Для вероятностных абсолютно непрерывных мер $\mathbb{P}_k \in \mathcal{P}_{ac}(\mathcal{X}_k)$, непрерывных функций стоимости $c_k(x,y) : \mathcal{X}_k \times \mathcal{Y} \mapsto \mathbb{R}$, $k \in \overline{K}$, и весов $\lambda_k \geq 0$, т. ч. $\sum_{k=1}^K \lambda_k = 1$, **двойственная задача SUOT барицентра** принимает следующий вид:

$$\mathcal{B}_u^* = \sup_{\substack{m\in\mathbb{R},\\ f_{[1:K]}\in\mathcal{C}^K(\mathcal{Y})\\ \sum_{k=1}^K \lambda_k f_k \equiv m}} \inf_{\gamma_k(\cdot|x_k)\in\mathcal{P}(\mathcal{Y})} \widetilde{\mathcal{L}}\left(f_{[1:K]}, \{\gamma_k(\cdot \mid x_k)\}_{[1:K]}, m\right), \tag{16}$$

где функционал $\widetilde{\mathcal{L}}\left(f_{[1:K]}, \{\gamma_k(\cdot \mid x_k)\}_{[1:K]}, m\right)$ определяется как

$$\widetilde{\mathcal{L}}\left(f_{[1:K]}, \{\gamma_k(\cdot \mid x_k)\}_{[1:K]}, m\right) \stackrel{\text{def}}{=}$$
$$\sum_{k=1}^K \lambda_k \left[\int_{\mathcal{X}_k} -\overline{\psi}_k \left(- \int_{\mathcal{Y}} \left(c_k(x_k,y) - f_k(y)\right) d\gamma_k(y \mid x_k) \right) d\mathbb{P}_k(x_k) + m \right].$$

Здесь и далее мы называем функции (потенциалы) $f_{[1:K]} : \mathcal{Y} \to \mathbb{R}$ *m-конгруэнтными* или удовлетворяющими условию *m-конгруэнтности*, если $\sum_{k=1}^K \lambda_k f_k \equiv m$ (для некоторого $m \in \mathbb{R}$).

В диссертации доказывается следующая теорема 4, устанавливающая связь между истинными планами γ^*, являющимися решениями задачи SUOT барицентра (6), и решениями предложенной задачи (16). В частности, теорема показывает, что левые маргиналы истинных планов $(\gamma_k^*)_x$ могут быть выражены через оптимальные потенциалы f_k^*, что

позволяет построить процедуру получения выборок данных из плотности маргинала на основе восстановленных потенциалов. Во-вторых, теорема гарантирует, что семейство оптимальных планов действительно является одним из решений задачи (16).

Теорема 4 (Решения задачи SUOT барицентра содержатся во множестве оптимальных седловых точек)**.** *Предположим, что существуют* m *и* $\{f_k^*(y)\in\mathcal{C}(\mathcal{Y})\}_{k=1}^K$*, доставляющие максимум в задаче* (16)*, при этом* $c_k(x,y)$ *– непрерывные функции стоимости, а* $\overline{\psi}_k$ *– непрерывно дифференцируемые функции. Пусть* $\{\gamma_k^*\}_{k=1}^K$ *– семейство SUOT планов между* $\mathbb{P}_k$ *и некоторым барицентром* $\mathbb{Q}^*$*. Тогда для каждого плана* γ_k^* *выполняется: (a) маргинал плана может быть представлен в виде:*

$$d(\gamma_k^*)_x(x) = \nabla\overline{\psi_k}(-(f_k^*)^{c_k}(x))d\mathbb{P}_k(x); \tag{17}$$

(b) соответствующий условный план $\gamma_k^*(\cdot|x_k)$ *удовлетворяет*

$$\gamma_k^*(\cdot|x_k) \in \arg\min_{\gamma_k(\cdot|x_k)} \widetilde{\mathcal{L}}(f_{[1:K]}^*, \{\gamma_k(\cdot|x_k))\}_{[1:K]}, m). \tag{18}$$

Теорема 4 показывает, что для некоторых оптимальных седловых точек $\{(f_k^*, \gamma_k^*)\}_{k=1}^K$ задачи (16) верно, что $\{\gamma_k^*\}_{k=1}^K$ является семейством истинных SUOT планов между $\mathbb{P}_{[1:K]}$ и $\mathbb{Q}^*$. В то же время, для любого семейства оптимальных $f_{[1:K]}^*$ множества $\arg\min_{\gamma_k(\cdot|x_k)}$ могут содержать не только истинные SUOT планы $\{\gamma_k^*\}_{k=1}^K$, но и другие функции, что соответствует известной проблеме *ложных решений* нейросетевых методов OT.

В диссертации также доказывается следующая теорема 5, устанавливающая связь между ошибкой решения разработанной задачи (16) с использованием тройки $(\widehat{m},\{\widehat{f}_k\}_{k=1}^K, \{\widehat{\gamma}_k\}_{k=1}^K)$ и ошибкой восстановленного решения $\{\widehat{\gamma}(\cdot|x_k)_k\}_{k=1}^K$ прямой задачи SUOT барицентра (6).

Теорема 5 (Оценки качества для восстановленных планов)**.** *Рассмотрим восстановленные* $\widehat{m}$*-конгруэнтные потенциалы* $\widehat{f}_{[1:K]}$*, такие что* $\widehat{f}_k \in \mathcal{C}(\mathcal{Y}), k \in \overline{K}$*, и условные планы* $\{\widehat{\gamma}_k(\cdot|x_k)\}_{k=1}^K, x_k \sim \mathbb{P}_k$*, такие что* $\widehat{\gamma}_k(\cdot|x_k)\in\mathcal{P}(\mathcal{Y}), k\in\overline{K}$*. Предположим, что выполняются условия теоремы 4, множество* $\mathcal{Y}$ *является выпуклым, а отображения* $y\mapsto c_k(x_k,y)-\widehat{f}_k(y)$ *являются* β*-сильно выпуклыми на* $\mathcal{Y}$*,* $k\in\overline{K}$*. Пусть также* $0<\eta \stackrel{def}{=} \nabla\overline{\psi}(-b)$*, где* b *– верхняя граница* $b = \max\limits_{k\in\overline{K}} \max\limits_{f_k\in\{\widehat{f}_k, f_k^*\}} \max\limits_{x_k\in\mathcal{X}_k, y\in\mathcal{Y}} \{c_k(x_k,y) - f_k(y)\}$*.*

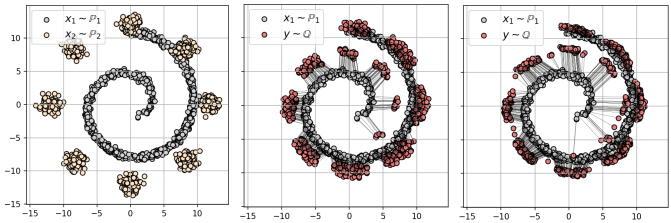
Введем разрывы двойственности:

$$\delta_1(\widehat{f}_{[1:K]}, \widehat{\gamma}_{[1:K]}) \stackrel{def}{=} \widetilde{\mathcal{L}}(\widehat{f}_{[1:K]}, \widehat{\gamma}_{[1:K]}, \widehat{m}) - \inf_{\gamma_k(\cdot|x_k)\in\mathcal{P}(\mathcal{Y})} \widetilde{\mathcal{L}}(\widehat{f}_{[1:K]}, \gamma_{[1:K]}, \widehat{m});$$

$$\delta_2(\widehat{f}_{[1:K]}) \stackrel{def}{=} \mathcal{B}_u^* - \inf_{\gamma_k(\cdot|x_k)\in\mathcal{P}(\mathcal{Y})} \widetilde{\mathcal{L}}(\widehat{f}_{[1:K]}, \gamma_{[1:K]}, \widehat{m}).$$

Тогда выполняется неравенство

$$\frac{2}{\eta\beta}\big(\delta_1+\delta_2\big) \geq \sum_{k=1}^K \lambda_k \int_{\mathcal{X}_k} \mathbb{W}_2^2\big(\widehat{\gamma}_k(\cdot|x), \gamma_k^*(\cdot|x)\big) d\mathbb{P}_k(x).$$



а) Входные распределения $\{\mathbb{P}_k\}$ б) Истинный SUOT барицентр $\mathbb{Q}^*$ в) Вычисленный SUOT барицентр $T_1 \# \mathbb{P}_1$

Рис. 3 — Барицентр распределения "Spiral" ($\mathbb{P}_1$) и смеси гауссовских распределений ($\mathbb{P}_2$), отображение переноса T_1 и распределение SUOT барицентра $\mathbb{Q}^* = T_1 \# \mathbb{P}_1$, полученные с помощью дискретного UOT (рассматривается как истинное решение) и предложенного метода в задаче *Spiral* $\rightarrow$ *Gaussian Mixture* при $\lambda_1 = \lambda_2 = \frac{1}{2}$, $\psi_1 = Id$, $\psi_2 = \mathrm{D}_{\mathrm{KL}}$.

Теорема 5 устанавливает оценки качества для восстановленных планов $\{\widehat{\gamma}_k(\cdot|x_k)\}_{k=1}^K$, исследуя *разрывы двойственности*, то есть ошибки при решении внутренней и внешней задач оптимизации в (16). Она показывает, что качество решения предложенной задачи оптимизации (16) ограничивает сверху качество аппроксимации барицентра $\mathbb{Q}^*$ с помощью восстановленных условных планов $\{\widehat{\gamma}(\cdot|x_k)_k\}_{k=1}^K$, примененных к распределениям $\mathbb{P}_k$.

В работе предлагается **вычислительный алгоритм** для оптимизации предложенного функционала (16), основанный на методах глубинного обучения. Условные планы рассматриваются как стохастические или детерминированные отображения и параметризуются в виде нейросетей T_{k,ω_k} с весами ω_k. Потенциалы параметризуются в виде нейросетей f_{k,θ_k} с параметрами θ_k, которые искусственно сконструированы таким образом, чтобы удовлетворять условию m-конгруэнтности. Параметры нейросетей оптимизируются с помощью процедуры градиентного подъема-спуска с использованием случайных выборок из входных распределений $\mathbb{P}_k$.

Для эмпирической оценки эффективности предложенного алгоритма приведены качественные и количественные оценки SUOT барицентров, полученных с его помощью, в следующих задачах: (a) вычисление SUOT барицентров двумерных распределений, страдающих от проблем дисбаланса классов и выбросов; (b) вычисление SUOT барицентров гауссовских распределений; (c) интерполяция распределений изображений на многообразии изображений. Результаты подтверждают превосходство предложенного алгоритма над существующими (сбалансированными) численными методами для барицентров.

Пример 2 (SUOT барицентр распределения "Spiral" и смеси гауссовских распределений)**.** *См. рис. 3.*

Четвертая глава посвящена задаче нахождения неполного ОТ отображения между двумя непрерывными вероятностными распределениями. Для решения данной задачи предлагается и теоретически обосновывается численный метод, основанный на использовании нейросетей. Далее представлены основные определения и результаты главы.

В данной главе рассматриваются компактные польские пространства $\mathcal{X}$, $\mathcal{Y}$ и непрерывная функция стоимости $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. В диссертации показано, что при таких предположениях IT планы $\pi^* \in \Pi^w(\mathbb{P},\mathbb{Q})$, доставляющие минимум в задаче IT (7), всегда существуют. Однако не гарантируется, что они детерминированы, то есть имеют вид $\pi^*=[\mathrm{id}_{\mathcal{X}},T^*]\sharp\mathbb{P}$, где T^* называется IT отображением.

В основе предложенного алгоритма для вычисления неполного транспорта (IT) лежит **max–min** реформулировка двойственной задачи IT (7). Для заданного веса $w \geq 1$ **двойственная IT задача** имеет вид:

$$\mathrm{Cost}_w(\mathbb{P},\mathbb{Q}) = \max_{f \leq 0} \min_{T:\mathcal{X}\to\mathcal{Y}} \mathcal{L}(f,T), \tag{19}$$

где функционал $\mathcal{L}(f,T)$ определяется как

$$\mathcal{L}(f,T) \stackrel{def}{=} \int_{\mathcal{X}} c\big(x,T(x)\big)d\mathbb{P}(x) - \int_{\mathcal{X}} f\big(T(x)\big)d\mathbb{P}(x) + w\int_{\mathcal{Y}} f(y)d\mathbb{Q}(y). \tag{20}$$

Здесь минимизация производится по измеримым отображениям $T : \mathcal{X} \to \mathcal{Y}$. Аналогично несбалансированному OT, функция f называется *потенциалом*.

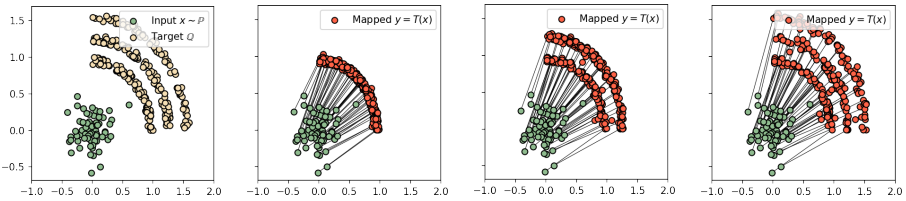
Выведенная двойственная IT задача (19) позволяет выполнять минимизацию по детерминированным функциям T. В то же время в диссертации доказывается следующая теорема, показывающая, что, решая задачу (19), можно получить истинные IT отображения T^*.

Теорема 6 (IT отображения содержатся во множестве оптимальных седловых точек)**.** *Пусть f^* – любой максимизатор в* (19)*. Если $\pi^* \in \Pi^w(\mathbb{P},\mathbb{Q})$ является детерминированным IT планом, то есть решает задачу* (7) *и имеет вид $\pi^* = [\mathrm{id}_{\mathcal{X}},T^*]\sharp\mathbb{P}$ для некоторого измеримого отображения $T^* : \mathcal{X} \to \mathcal{Y}$, то $T^* \in \arg\min_{T:\mathcal{X}\to\mathcal{Y}} \mathcal{L}(f^*,T)$.*

В диссертации предлагается **вычислительный алгоритм** для оптимизации предложенного функционала (20) с использованием методов глубинного обучения. IT отображение T и потенциал f параметризуются как нейросети T_θ и f_ψ с параметрами θ и ψ соответственно. Чтобы сделать f_ψ неположительной функцией, последний слой задается как $x \mapsto -|x|$.

Для эмпирической оценки эффективности алгоритма проведен ряд вычислительных экспериментов: (а) эксперименты с двумерными распределениями, демонстрирующие способность алгоритма находить отображения между входным распределением и заданной частью целевого; (б) применение алгоритма к задаче преобразования изображений.

Пример 3 (IT отображение между гауссовским распределением и распределением «Archs»)**.** *См. рис. 4.*



а) Входное и целевое распределения. б) IT отображение, $w = 3$. в) IT отображение, $w = \frac{3}{2}$. г) IT отображение, $w = 1$.

Рис. 4 — Восстановленные IT отображения между гауссовским распределением и «Archs» при различных w. Изменение w позволяет находить отображения в заданные части $(1/w)$ целевого распределения.

Заключение обобщает основные результаты работы и обсуждает возможные направления дальнейших исследований в данной области.

Публикации автора по теме диссертации

В изданиях из списка ВАК РФ и приравненных к ним

1 ***Gazdieva, M.***, *Mokrov, P., Rout, L., Korotin, A., Kravchenko, A., Filippov, A., & Burnaev, E.* An optimal transport perspective on unpaired image super-resolution // Journal of Optimization Theory and Applications. — 2025. — Vol. 207. — Art. 40. **Scopus (Q1)**.

2 ***Газдиева, М.***, *Коротин, А.* Обзор минимаксных параметрических методов вычисления оптимальных транспортных отображений. // Труды Института Системного Анализа Российской Академии Наук. — 2026. — Т. 76, № 3. — С. 16-23. **ВАК (К2)**.

В сборниках трудов конференций

3 ***Gazdieva, M.***, *Asadulaev, A., Burnaev, E., & Korotin, A.* Light unbalanced optimal transport // Advances in Neural Information Processing Systems. — 2024. — Vol. 37. — P. 93907-93938. **Scopus (A*)**.

4 ***Gazdieva, M.***, *Choi, J., Kolesov, A., Choi, J., Mokrov, P., & Korotin, A.* Robust Barycenter Estimation using Semi-Unbalanced Neural Optimal Transport // International Conference on Learning Representations. — 2025. [Electronic resource] URL: https://openreview.net/pdf?id=CI5Cj0vktS (date of reference 20.08.2025). **Scopus (A*)**.

5 ***Gazdieva, M.***, *Korotin, A., Selikhanovych, D., & Burnaev, E.* Extremal domain translation with neural optimal transport // Advances in Neural Information Processing Systems. — 2023. — Vol. 36. — P. 40381-40413. **Scopus (A*)**.