Федеральное государственное автономное образовательное
учреждение высшего образования
“Национальный исследовательский университет
“Высшая школа экономики”

Факультет математики

На правах рукописи

Пучкин Никита Андреевич

Построение оптимальных оценок с помощью метода адаптивных весов в задачах обучения с размеченными и неразмеченными данными

Резюме диссертации
на соискание ученой степени
кандидата математических наук

Научный руководитель:
к.ф.-м.н.
Спокойный Владимир Григорьевич

Москва – 2023

Содержание

Введение

Задача распознавания закономерностей в больших массивах данных привлекла к себе пристальное внимание в последнее время. Задача статистического обучения находит все более приложений в эру больших данных. Спектр областей применения методов машинного обучения поражает своей широтой: от биоинформатики (например, предсказание структуры белка [1]) и финансов (предсказание кредитного рейтинга [2] и вероятности банкротства [3]) до распознавания речи [4], анализа изображений [5, 6] и их восстановления [7, 8]. Рост интереса к машинному обучению привел к существенному прогрессу в теоретическом обосновании и понимании принципов работы существующих методов, а также подтолкнул исследователей к разработке новых, более эффективных алгоритмов. Обычно анализ методов машинного обучения опирается на результаты из теории вероятностей и математической статистики. Как правило, предполагается наличие некоторой вероятностной модели, из которой сгенерированы наблюдаемые данные. Исходя из нее и предполагается сделать статистический вывод.

Задачи статистического обучения можно разбить на 3 большие группы: обучение с учителем, частичное обучение и обучение без учителя. В задаче обучения с учителем каждое из наблюдений имеет некоторую метку, также называемую откликом или целевой переменной (как правило, это действительное число). Цель статистика состоит в предсказании метки у вновь сгенерированных наблюдений, не представленных в выборке. Наиболее популярными примерами задач обучения с учителем являются классификация и регрессия. Напротив, в задаче обучения без учителя поиск закономерностей осуществляется на основе неразмеченных наблюдений. Примеры таких задач включают в себя кластеризацию и восстановление гладких многообразий. Наконец, в задачах частичного обучения предполагается наличие небольшой порции размеченных данных и большого количества неразмеченных.

Цели данной диссертации состоят в разработке новых адаптивных алгоритмов для задач обучения с учителем и без учителя и доказательстве строгих теоретических оценок на качество их работы. Глава 1 посвящена многоклассовой классификации. Предложен адаптивный метод ближайших соседей, Алгоритм 1. Несмотря на то, что специалисты по математической статистике давно знакомы с методом k ближайших соседей, неасимптотический анализ данного метода был проведен сравнительно недавно. В [9] автором доказана минимаксно оптимальная верхняя оценка на избыточный риск метода взвешенных ближайших соседей. К сожалению, результаты работы [9] получены при достаточно ограничительных предположениях на распределение данных. В частности, предполагается, что носителем распределения векторов признаков является компактное множество, и плотность распределения на нем отделена от нуля. Частичное решение этой проблемы было предложено в работах [10] и [11], где авторы представили условие, связывающее распределения целевых переменных и наборов признаков. Важные шаги в понимании границ применимости метода k ближайших соседей были предприняты в работах [12] и [13]. В частности, авторы статьи [12]доказали, что метод ближайших соседей с универсальным выбором параметра k для всех тестовых точек является неоптимальным при довольно естественных предположениях. В работе [13] было предложено использовать неразмеченные данные для оптимального выбора значения k для каждой тестовой точки. К сожалению, данных подход не всегда применим, поскольку неразмеченные данные могут быть недоступны. Данная диссертация развивает идею пошаговой агрегации [3]. Предложен метод адаптивных весов, основанный на комбинации оценок метода ближайших соседей. Хотя обобщение алгоритма с бинарной классификации на многоклассовую достаточно очевидно, статистический анализ предложенного метода оказался гораздо более трудоемким. Это связано с тем, что в диссертации накладываются более слабые условия на распределение данных по сравнению с [3]. Показано, что при тех же предположениях, что и в работе [12], представленный в диссертации алгоритм достигает оптимальных порядков сходимости с точностью до логарифмического множителя. Заметим, что, в отли-

чие от [13], процедура не требует дополнительной неразмеченной выборкой. Классификатор с такими свойствами представлен в литературе по машинному обучению и математической статистике впервые.

Глава 2 посвящена задаче восстановления гладкого многообразия низкой размерности в $\mathbb{R}^D$ по неточным наблюдениям, получившей развитие в ряде исследований. К сожалению, существующие методы оценки многообразий либо рассчитаны на случай малой амплитуды шума (например, [8, 14–16]), либо предполагают распределение шума известным (см., например, [17, 18]). В данной диссертации рассмотрена новая постановка задачи, ранее не встречавшаяся в литературе по указанной теме. А именно, в отличие от [17, 18], не предполагается жестких условий на распределение шума. В то же время допустимая амплитуда шума значительно выше, чем в работах [8, 14–16, 19]. Хотя накладываемые предположения вполне реалистичны, они значительно отличаются от обычно рассматриваемых в литературе условий. Поэтому неудивительно, что существующие алгоритмы либо не имеют теоретических верхних оценок на точность восстановления многообразия, либо неоптимальны. Таким образом, была поставлена амбициозная задача построения оптимальных оценок многообразия в описанной постановке. В данной диссертации развивается идея структурной адаптации [20, 21]. Предложен новый алгоритм (Алгоритм 2) для восстановления проекций наблюдений на скрытое многообразие. Благодаря Алгоритму 2 построена состоятельная оценка скрытого многообразия (см. Теорему 4). Также доказана новая минимаксная нижняя оценка на точность восстановления гладкого многообразия по неточным наблюдениям (Теорема 5), подтверждающая оптимальность предложенной процедуры.

Задачи, рассмотренные в Главах 1 и 2 существенно отличаются друг от друга, но предложенные методы, Алгоритм 1 и Алгоритм 2, основаны на схожих идеях. Основным элементом алгоритмов является так называемый метод адаптивных весов, правильно приспособленный под задачи многоклассовой классификации и восстановления многообразия. Адаптивность предложенных процедур (то есть способность неявно выполнять автоматический частичный подбор параметров, упрощая настройку алгоритмов) увеличивает их практическую значимость. Также стоит отметить, что Алгоритмы 1 и 2 можно скомбинировать и применить в задачах частичного обучения.

Положения диссертации, выносимые на защиту, состоят в следующем.

1. Предложен алгоритм многоклассовой классификации, Алгоритм 1, основанный на агрегации оценок метода k ближайших соседей. Процедура автоматически выбирает близкое к оптимальному значение k для каждой тестовой точки и каждого класса. Помимо этого, алгоритм адаптируется к гладкости целевой функции.

2. Получена оценка больших уклонений избыточного риска классификатора, выдаваемого Алгоритмом 1, при мягких предположениях на распределение данных. Полученные теоретические результаты ранее не встречались в литературе по данной теме и влекут оптимальность предложенного метода.

3. Предложен новый алгоритм оценки проекций неточных наблюдений на скрытое многообразие, Алгоритм 2, основанный на идее структурной адаптации.

4. Благодаря анализу Алгоритма 2 доказана новая верхняя оценка на точность восстановления гладкого многообразия низкой размерности по неточным наблюдениям.

5. Доказана новая минимаксная нижняя оценка на точность восстановления гладкого многообразия по неточным наблюдениям, из которой следует оптимальность предложенного метода.

Результаты диссертации были представлены на следующих конференциях, школах и семинарах.

1. Мини-конференция “New frontiers in high-dimensional probability and statistics” (“Новые рубежи в теории вероятностей и математической статистике”), Москва, 23–24 февраля 2018 г. Тема доклада: “Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification” (“Агрегация локальных оценок методом поточечной адаптации в задаче многоклассовой классификации”).

2. Двенадцатая международная вильнюсская конференция по теории вероятностей и математической статистике и Ежегодная встреча Института математической статистики (IMS) по теории вероятностей и математической статистике 2018, Вильнюс, Литва, 2–6 июля 2018 г. (постер). Тема постера: “Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification” (“Агрегация локальных оценок методом поточечной адаптации в задаче многоклассовой классификации”).

3. Исследовательский семинар “Структурное обучение”, Москва, 11 октября 2018 г. Тема доклада: “Manifold Learning” (“Восстановление многообразий”).

4. Зимняя школа “New frontiers in high-dimensional probability and statistics 2” (“Новые рубежи в теории вероятностей и математической статистике 2”), Москва, 22–23 февраля 2019 г. Тема доклада: “Manifold estimation from noisy observations” (“Оценка многообразия по неточным наблюдениям”).

5. 49-я летняя школа в Сен-Флуре, Сен-Флур, Франция, 7–19 июля 2019 г. Тема доклада: “Manifold estimation from noisy observations” (“Оценка многообразия по неточным наблюдениям”).

6. Конференция “Structural Inference in High-Dimensional Models 2” (“Структурный вывод в многомерных моделях 2”), Пушкин, Санкт-Петербург, 26–30 августа 2019 г. (постер). Тема постера: “Structure-adaptive manifold estimation” (“Структурно-адаптивная оценка многообразий”).

7. Осенняя школа НИУ ВШЭ и Яндекса по генеративным моделям, Москва, 26–29 ноября 2019 г. Тема доклада: “Structure-adaptive manifold estimation” (“Структурно-адаптивная оценка многообразий”).

8. Доклад на исследовательском семинаре “Структурное обучение”, Москва, 3 декабря 2019 г. “Sample complexity of learning a manifold with an unknown dimension” (“Выборочная сложность оценки многообразия неизвестной размерности”).

9. Конференция “Mathematical Methods of Statistics” (“Математические методы статистики”), Люмини, Франция, 16–20 декабря 2019 г. Тема доклада: “Structure-adaptive manifold estimation” (“Структурно-адаптивная оценка многообразий”).

10. Конференции факультета компьютерных наук НИУ ВШЭ по машинному обучению, трек фундаментальных исследований, 18–20 ноября 2020 г. Тема доклада: “Структурно-адаптивная оценка многообразий”.

Основные результаты диссертации были опубликованы в двух статьях в рецензируемых научных изданиях: “Адаптивный метод k ближайших соседей для задачи многоклассовой классификации” (ESAIM: Probability & Statistics, [22]) и “Структурно-адаптивная оценка многообразий” (Journal of Machine Learning Research, [23]).

Содержание диссертации и представленные результаты отражают персональный вклад автора. Результаты исследований были подготовлены к публикации в сотрудничестве с научным руководителем, причем вклад автора диссертации был решающим. Представленные результаты получены автором лично.

Диссертация состоит из введения, двух глав, заключения и списка литературы. В начале каждой главы присутствует обзор литературы на релевантную тему. Объем диссертации составляет 100 страниц, включая 95 страниц текста, 5 таблиц и 5 рисунков. Список литературы занимает 5 страниц и включает в себя 83 наименования.

Глава 1
Многоклассовая классификация

1.1. Постановка задачи

Задача многоклассовой классификации является естественным обобщением классической задачи классификации. Это задача обучения с учителем, в которой по данной обучающей выборке $S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, где $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$, $Y_i \in \mathcal{Y} = \{1, \ldots, M\}$, $1 \leqslant i \leqslant n$, $M > 2$, требуется построить решающее правило f для предсказания метки Y тестовой точки X. Классическая постановка задач статистического обучения предполагает, что элементы обучающей выборки (X_i, Y_i) равно как и тестовая пара (X, Y) сгенерированы независимо из некоторого неизвестного распределения $\mathcal{D}$ на $\mathcal{X} \times \mathcal{Y}$. Функционалом качества решающего правила f является вероятность неправильной классификации тестовой пары (X, Y):

$$R(f) = \mathbb{P}_{(X,Y)\sim\mathcal{D}}\left(Y \neq f(X)\right).$$

Цель обучения состоит в построении классификатора $\widehat{f}$, для которого $R(\widehat{f})$ мал, насколько это возможно. На практике нередко встречаются ситуации, когда необходимо отнести объект к одному из нескольких классов. Примерами могут служить предсказание структуры белка [1] или классификация отрезков ДНК [24] в задачах биоинформатики, предсказание кредитного рейтинга [2] в финансовой отрасли, классификация объекта на изображении [5], распознавание речи [4] и др.

1.2. Непараметрическая многоклассовая классификация и обзор литературы

Обозначим через $\mathbb{P}_X$ маргинальное распределение X и предположим, что оно абсолютно непрерывно относительно меры Лебега μ с плотностью $p(X)$. Для каждого $m \in \{1, \ldots, M\}$ положим $\eta_m(X)$ равной условной вероятности $\mathbb{P}(Y = m|X)$. В рассматриваемой модели наилучшим классификатором является байесовское решающее правило f^*, определяемое по формуле

$$f^*(X) = \operatorname*{argmax}_{1\leqslant m\leqslant M} \eta_m(X). \tag{1.1}$$

К сожалению, значения функций $\eta_1(X), \ldots, \eta_M(X)$ неизвестны, так как неизвестно распределение $\mathcal{D}$, однако их можно оценить.

Также стоит отметить, что так как для любого классификатора f выполнено неравенство $R(f) \geqslant R(f^*)$, то разумно характеризовать качество предсказаний классификатора f его избыточным риском

$$\mathcal{E}(f) = R(f) - R(f^*)$$

по отношению к байесовскому классификатору f^*. Одним из наиболее популярных и широко используемых методов оценки функций $\eta_1(X), \ldots, \eta_M(X)$ является метод (взвешенных) k ближайших соседей. Как только для тестовой точки $X \in \mathcal{X}$ Метод построены оценки $\widehat{\eta}_1^{(NN)}(X), \ldots, \widehat{\eta}_M^{(NN)}(X)$ значений $\eta_1(X), \ldots, \eta_M(X)$ по правилу k ближайших соседей, классификатор предсказывает метку

$$\widehat{f}^{(NN)}(X) = \operatorname*{argmax}_{1\leqslant m\leqslant M} \widehat{\eta}_m^{(NN)}(X).$$

Несмотря на простоту и известность метода, несколько новых неасимптотических результатов были получены сравнительно недавно. В работе [9] автор рассматривает оценку взвешенных ближайших соседей с гладкой функцией $\eta_1(x)$ и получает формулу оптимального вектора неотрицательных весов. Более того, автор показывает, что возможно получение более быстрых порядков сходимости, если позволить весам принимать отрицательные значения. Однако в [9] предполагается, что плотность маргинального распределения векторов признаков отделена от нуля на своем носителе. Это ограничительное условие ослабляется в работах, [10] и [11]. В статье [10] авторы ввели новое условие гладкости $\eta_1(x)$, адаптированное под метод ближайших соседей. В частности, оно допускает, что признаки могут быть неограниченными. В то же время, у этого условия имеется свой недостаток, состоящий в том, что задано неявное соотношение между распределением признаков и меток. Этого недостатка лишены условие минимальной массы и условие на «хвосты» распределения $\mathbb{P}_X$, предложенные в [12]. Данные условия допускают, что плотность $\mathbb{P}_X$ может быть сколь угодно близка к нулю. Заметим, что качество оценки $\widehat{\eta}_m^{(NN)}(X)$, а следовательно, и классификатора $\widehat{f}^{(NN)}$, сильно зависит от выбора параметра k. Более того, как было указано в работах [13] и [12], универсальный выбор этого параметра ведет к неоптимальным оценкам, в то время как индивидуальный выбор k для каждой тестовой точки показывает лучший результат. В случае многоклассовой классификации данная проблема обостряется еще сильнее, так как оптимальное значение k может быть разным для каждого класса, что существенно усложняет настройку параметра. Чтобы избавиться от этой проблемы, можно рассмотреть последовательность натуральных чисел $n_1, \ldots, n_K$, для каждого из них вычислить оценки метода взвешенных ближайших соседей и использовать классификатор, основанный на выпуклой комбинации этих оценок.

Агрегация оценок метода ближайших соседей – ключевой шаг предложенного алгоритма. Для этой цели используется процедура пошаговой агрегации (SSA), описанная в [3], где рассматривалась задача бинарной классификации. Достоинством метода является то, что, в отличие от экспоненциального взвешивания [25–27], зеркального усреднения [28, 29], минимизации эмпирического риска [30] и Q-агрегации [31, 32], действующих *глобально*, SSA подбирает вес каждого классификатора в зависимости от тестовой точки X. Недостатком алгоритма, предложенного в [3], являются ограничения, накладываемые в связи с использованием дивергенции Кульбака-Лейблера, которые, хоть и естественны для некоторых задач (см., например, [33, 34]), являются абсолютно лишними в задаче классификации. В данной главе показано, что можно получить аналогичные результаты при более слабых предположениях.

Наконец, стоит отметить, что скорость сходимости непараметрических оценок часто оказывается медленной, если размерность d велика. Как было показано в [35], а затем и в [36], в задачах непараметрической классификации (как бинарной, так и многоклассовой) можно достичь быстрых порядков сходимости при определенных условиях. В данной диссертации будет использован аналогичный подход.

1.3. Научный вклад

Основными результатами данной главы являются:

- вычислительно эффективный алгоритм многоклассовой классификации, основанный на агрегации оценок метода ближайших соседей, который
  - (а) автоматически выбирает близкое к оптимальному число соседей для каждой тестовой точки и каждого класса;
  - (б) адаптируется к неизвестной гладкости функций $\eta_1(\cdot), \ldots, \eta_M(\cdot)$;

- теоретические верхние оценки на избыточный риск предложенного классификатора при мягких предположениях. Оценки подтверждают, что построенный метод является оптимальным в минимаксном смысле с точностью до логарифмического множителя.

1.4. Обозначения и модельные предположения

Начнем с простого наблюдения. Обозначим

$$\varphi(t) = \left(\frac{1}{2M} \vee t\right) \wedge \left(1 - \frac{1}{2M}\right). \tag{1.2}$$

Нетрудно заметить, что композиция

$$\theta_m(X) = \varphi(\eta_m(X)) \equiv \left(\frac{1}{2M} \vee \eta_m(X)\right) \wedge \left(1 - \frac{1}{2M}\right),$$

Удовлетворяет неравенству

$$f^*(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \eta_m(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \theta_m(X),$$

где, как и прежде, f^* – байесовский классификатор. Таким образом, вместо $\eta_m(x)$, можно оценить значение $\theta_m(x)$ в точке x и рассмотреть классификатор

$$\widehat{f}(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \widehat{\theta}_m(X), \tag{1.3}$$

где $\widehat{\theta}_m(x)$ – оценка $\theta_m(x)$, $1 \leqslant m \leqslant M$, в точке x. Задача свелась к построению оценок значений $\theta_m(x)$, $1 \leqslant m \leqslant M$.

Зафиксируем класс m и трансформируем метки в бинарные: $\mathbb{1}(Y_i = m)$. Очевидно, что

$$\left(\mathbb{1}(Y_i = m) \,|X_i\right) \sim \text{Bernoulli}(\eta_m(X_i)).$$

В работе рассматривается взвешенный аналог метода метода k ближайших соседей, то есть строятся оценки вида $\widetilde{\theta}_m^w(x) = \varphi(\widetilde{\eta}_m^w(x))$, где

$$\widetilde{\eta}_m^w(x) = \frac{\sum\limits_{i=1}^n w_i(X_i, x)\mathbb{1}(Y_i = m)}{\sum\limits_{i=1}^n w_i(X_i, x)} \equiv \frac{S_m^w(x)}{N_w(x)}. \tag{1.4}$$

Для краткости введены обозначения $S_m^w(x) = \sum\limits_{i=1}^n w_i(X_i, x)\mathbb{1}(Y_i = m)$, $N_w(x) = \sum\limits_{i=1}^n w_i(X_i, x)$. Неотрицательные веса $w_i(X_i, x)$ зависят от расстояния X_i между x и отличны от нуля только в том случае, когда X_i попадает в число k ближайших соседей к x; иначе, $w_i(X_i, x) = 0$. В данной главе рассматриваются веса вида

$$w_i = w_i(X_i, x) = \mathcal{K}\left(\frac{\|X_i - x\|}{h}\right), \tag{1.5}$$

где $h = h(\mathsf{k})$ – расстояние до k-го ближайшего соседа точки X. Функция K называется локализующим ядром. В данной главе предполагается, что эта функция удовлетворяет следующим условиям:

- $\mathcal{K}(t)$ является невозрастающей функцией на $[0, +\infty)$,
- $\mathcal{K}(0) = 1$,
- $\mathcal{K}(1) \geqslant \dfrac{1}{2}$,
- $\mathcal{K}(t) = 0, \quad \forall\, t > 1.$

(A1)

Приведем несколько примеров ядер, удовлетворяющих условию (A1). Во-первых, прямоугольное ядро $\mathcal{K}(t) = \mathbb{1}\,(0 \leqslant t \leqslant 1)$ удовлетворяет (A1), и следовательно, результаты данной главы распространяются и на классический (невзвешенный) метод k ближайших соседей. Во-вторых, несложно проверить, что ядра $\mathcal{K}(t) = (1 - t^2/2)\mathbb{1}(0 \leqslant t \leqslant 1)$ и $\mathcal{K}(t) = e^{-t^2/2}\mathbb{1}(0 \leqslant t \leqslant 1)$ также удовлетворяют этому условию. Отметим, что без ограничения общности предполагается, что с вероятностью 1 k-ый ближайший сосед определен однозначно. Если имеется несколько кандидатов на место k-ого ближайшего соседа, то можно использовать процедуру рандомизации, описанную в [10].

Качество метода k ближайших соседей сильно зависит от выбора параметра k. Задача сильно усложняется тем, что для каждого класса m и для каждой тестовой точки X оптимальное значение k может быть разным. Поэтому вместо использования одного универсального значения k, зафиксируем возрастающую последовательность натуральных чисел $\{n_k : 1 \leqslant k \leqslant K\}$, удовлетворяющую следующему свойству: существуют константы $0 < u_0 < u < 1$ такие, что

$$n_1 \leqslant a, \quad n_K \geqslant bn^{2/(d+2)}, \quad \text{and} \quad 2u_0 \leqslant \frac{n_{k-1}}{n_k} \leqslant \frac{u}{2}, \quad \text{for all } 1 \leqslant k \leqslant K. \tag{A2}$$

Кадому n_k соответствует набор весов $w_1^{(k)}, \ldots, w_n^{(k)}$, определяемый соотношением

$$w_i^{(k)} = w_i^{(k)}(X_i, x) = \mathcal{K}\left(\frac{\|X_i - x\|}{h_k}\right), \tag{1.6}$$

где h_k – расстояние до n_k-ого ближайшего соседа, а также оценка взвешенного метода n_k ближайших соседей:

$$\widetilde{\theta}_m^{(k)}(x) = \varphi\left(\widetilde{\eta}_m^{(k)}(x)\right) \equiv \left(\frac{1}{2M} \vee \widetilde{\eta}_m^{(k)}(x)\right) \wedge \left(1 - \frac{1}{2M}\right), \tag{1.7}$$

$$\widetilde{\eta}_m^{(k)}(x) = \frac{S_m^{(k)}(x)}{N_k(x)}, \tag{1.8}$$

где $S_m^{(k)}(x) = \sum\limits_{i=1}^{n} w_i^{(k)}(X_i, x)\mathbb{1}(Y_i = m)$, $N_k(x) = \sum\limits_{i=1}^{n} w_i^{(k)}(X_i, x)$. Далее с помощью алгоритма из работы [3] конструируются агрегированные оценки $\widehat{\theta}_1(x), \ldots, \widehat{\theta}_M(x)$, после чего предсказание метки в точке x осуществляется по формуле (1.3). Более детальное описание алгоритма, который будем называть MSSA (сокр. multiclass spatial stagewise aggregation), приведено в Главе 1.5.

Для доказательства состоятельности предложенного алгоритма будут получены верхние оценки на обобщающую способность $\mathbb{P}_{(X,Y)\sim\mathcal{D}}\left(Y \neq \widehat{f}(X)\middle|S_n\right)$ классификатора $\widehat{f}$, которые будут выполнены в среднем или с большой вероятностью по всем обучающим выборкам S_n. В качестве промежуточного шага будет получены оценки на величину $\max\limits_{1\leqslant m\leqslant M} |\widehat{\theta}_m(x) - \theta_m^*(x)|$, а также качество метода взвешенных ближайших соседей в фиксированной точке x при мягких предположениях. А именно, помимо (A1) и (A2), предполагается следующее. Во-первых, функции $\eta_m(\cdot)$ (L, α)-Гельдеровы, т.е. существуют константы $L > 0$ и $\alpha > 0$ такие, что для всех $x, x' \in \mathcal{X}$ и $1 \leqslant m \leqslant M$ выполнено неравенство

$$|\eta_m(x) - \eta_m(x')| \leqslant L\|x - x'\|^\alpha. \tag{A3}$$

Во-вторых, предполагается наличие условия малого шума, чтобы избежать проклятия размерности, которое часто появляется в задаче непараметрической классификации при больших d. В диссертации используется обобщение условия малого шума Маммена-Цыбакова [37]

на многоклассовый случай. А именно, предполагается, что существуют константы $B > 0$ и $\beta \geqslant 0$ такие, что для любого $t > 0$ выполнено

$$\mathbb{P}\left(\eta_{(1)}(X) - \eta_{(2)}(X) < t\right) \leqslant Bt^{\beta} \tag{A4}$$

Здесь и далее, для каждого x $\eta_{(1)}(x) \geqslant \ldots \geqslant \eta_{(M)}(x)$ обозначают упорядоченный по убыванию набор значений $\eta_1(x), \ldots, \eta_M(x)$.

Для статистического анализа потребуются еще два условия: условие минимальной массы и условие на «хвосты» маргинального распределения $\mathbb{P}_X$ (см. [12]) случайного вектора X. Первое условие подразумевает, что существуют положительные константы $\varkappa$ и r_0 такие, что для всех $r \in (0, r_0]$ и $x \in \mathrm{supp}(\mathbb{P}_X)$ выполнено неравенство

$$\mathbb{P}(X \in B(x, r)) \geqslant \varkappa p(x) r^d, \tag{A5}$$

где $B(x, r)$ – Евклидов шар радиуса r с центром в точке x и $p(x)$ – плотность меры $\mathbb{P}_X$. Условие на «хвосты» маргинального распределения $\mathbb{P}_X$ подразумевает существование положительных констант C, ε_0 и p таких, что для любого $\varepsilon \in (0, \varepsilon_0]$ выполнено

$$\mathbb{P}\left(p(X) < \varepsilon\right) \leqslant C\varepsilon^p. \tag{A6}$$

В работе [12] (Теорема 4.1) показано, что условия (A5) и (A6) необходимы для анализа классификаторов и не могут быть опущены. В отличие от условий в статье [9], предположение (A6) допускает, что носитель $\mathbb{P}_X$ может быть неограниченным. Нетрудно доказать, что если $\mathbb{E}\|X\|^r < \infty$ для некоторого $r > 0$, то X удовлетворяет условию (A6) с $p = r/(r + d)$. В дальнейшем предполагается, что параметр p из условия (A6) больше $\alpha/(2\alpha + d)$.

1.5. Метод адаптивных весов для многоклассовой классификации

В данной главе представлен адаптивный алгоритм многоклассовой классификации (Алгоритм 1). Процедура получает на вход возрастающую последовательность натуральных чисел $\{n_k : 1 \leqslant k \leqslant K\}$, удовлетворяющую (A2), обучающую выборку $S_n = \{(X_i, Y_i) : 1 \leqslant i \leqslant n\}$, тестовую точку $X \in \mathcal{X}$ и набор положительных чисел $\{z_k : 1 \leqslant k \leqslant K\}$. Числа $z_1, \ldots, z_K$ называются критическими значениями.

Заметим, что, по построению, $\widetilde{\theta}_m^{(k)}(x) \in [1/(2M), 1 - 1/(2M)]$, и поэтому оценка $\widehat{\theta}_m^{(k)}(x)$ также принадлежит отрезку $[1/(2M), 1 - 1/(2M)]$ и величина $\mathcal{K}\left(\widetilde{\theta}_m^{(k)}(x), \widehat{\theta}_m^{(k-1)}(x)\right)$, которая является расстоянием Кульбака-Лейблера между двумя распределениями Бернулли с параметрами $\widetilde{\theta}_m^{(k)}(x)$ и $\widehat{\theta}_m^{(k-1)}(x)$, определена корректно.

Оценим время работы Алгоритма 1. Из условия (A2) следует, что $K = O(\log n)$ и, следовательно, требуется $O\left(Mn \log n\right)$ операций, чтобы найти оценки по методу ближайших соседей для всех классов и $O(\log n)$ операций, чтобы их сагрегировать. Таким образом, для предсказания метки одной тестовой точки Алгоритм 1 требует $O\left(Mn \log n\right)$ операций. Если необходимо предсказать метки нескольких тестовых точек, вычисления можно проводить параллельно.

1.6. Теоретические свойства алгоритма

1.6.1. Главный результат

Теорема 1. *Пусть выполнены условия* (A1) – (A5), *причем в условии* (A6) $p > \alpha/(2\alpha + d)$. *Выберем параметры* $z_1, \ldots, z_K$ *согласно формуле*

$$z_k = \frac{8M^2}{u_0} \log \frac{12KM}{\delta_*}, \quad 1 \leqslant k \leqslant K, \tag{1.9}$$

Алгоритм 1 MSSA

Исходные данные: последовательность натуральных чисел $\{n_k : 1 \leqslant k \leqslant K\}$, удовлетворяющая (A2), набор действительных чисел $\{z_k : 1 \leqslant k \leqslant K\}$, обучающая выборка $S_n = \{(X_i, Y_i) : 1 \leqslant i \leqslant n\}$, тестовая точка $x \in \mathcal{X}$.

Цикл m **от** 1 **до** M **:**

  Для каждого k от 1 до K вычислить веса $w_i^{(k)} = w_i^{(k)}(X_i, x)$, $1 \leqslant i \leqslant n$,

  по формуле (1.6), где ядро $\mathcal{K}$ удовлетворяет условию (A1), и вычислить $\widetilde{\theta}_m^{(k)}(x)$

  по формулам (1.7) и (1.8).

  Положить $\widehat{\theta}_m^{(1)}(x) = \widetilde{\theta}_m^{(1)}(x)$.

  Цикл k **от** 2 **до** K **:**

    Вычислить $N_k(x) = \sum\limits_{i=1}^{n} w_i^{(k)}(X_i, x)$ и

$$\mathcal{K}\left(\widetilde{\theta}_m^{(k)}(x), \widehat{\theta}_m^{(k-1)}(x)\right) = \widetilde{\theta}_m^{(k)}(x) \log \frac{\widetilde{\theta}_m^{(k)}(x)}{\widehat{\theta}_m^{(k-1)}(x)} + \left(1 - \widetilde{\theta}_m^{(k)}(x)\right) \log \frac{1 - \widetilde{\theta}_m^{(k)}(x)}{1 - \widehat{\theta}_m^{(k-1)}(x)}.$$

    Вычислить $\gamma_k = \mathbb{1}\left(N_k(x)\mathcal{K}\left(\widetilde{\theta}_m^{(k)}(x), \widehat{\theta}_m^{(k-1)}(x)\right) \leqslant z_k\right)$.

    Обновить оценку $\widehat{\theta}_m^{(k)}(x) = \gamma_k \widetilde{\theta}_m^{(k)}(x) + (1 - \gamma_k)\widehat{\theta}_m^{(k-1)}(x)$.

  Конец цикла

  Взять финальную оценку $\widehat{\theta}_m(x) = \widehat{\theta}_m^{(K)}(x)$.

Конец цикла

Результат: предсказанная метка $\widehat{f}(x) = \underset{1 \leqslant m \leqslant m}{\operatorname{argmax}} \left\{\widehat{\theta}_m(x)\right\}$.

причем

$$\delta_* = \begin{cases} \left(\frac{M^3 \log n}{np_0}\right)^{\frac{\alpha(2+\beta)}{2\alpha+d}}, & \text{если } \exists\, p_0 : p(x) \geqslant p_0 \,\forall\, x \in supp(\mathbb{P}_X), \\ \psi_*^{r_*}, & \text{иначе,} \end{cases} \tag{1.10}$$

где $r_* = \log \psi_*^{-1}$ *и*

$$\psi_* = \left(\frac{M^3 \log^2 n}{n}\right)^{\frac{\alpha}{\alpha\beta/p+2\alpha+d}}.$$

Пусть $\widehat{\theta}_1(\cdot), \ldots, \widehat{\theta}_M(\cdot)$ *– оценки, полученные с помощью Алгоритма 1. Тогда, если объем выборки* n *достаточно большой, для избыточного риска классификатора* $\widehat{f}(X) = \underset{1 \leqslant m \leqslant M}{\operatorname{argmax}} \widehat{\theta}_m(X)$ *выполнена оценка*

$$\mathbb{E}_{S_n}\mathcal{E}(\widehat{f}) \lesssim \begin{cases} \left(\frac{M^3 \log n}{np_0}\right)^{\frac{\alpha(1+\beta)}{2\alpha+d}}, & \text{если } \exists\, p_0 : p(x) \geqslant p_0 \,\forall\, x \in supp(\mathbb{P}_X), \\ \left(\frac{M^3 \log^2 n}{n}\right)^{\frac{\alpha(1+\beta)}{\alpha\beta/p+2\alpha+d}}, & \text{иначе.} \end{cases} \tag{1.11}$$

Более того, для любого $\delta \in (0, 1)$*, если*

$$z_k = \frac{8M^2}{u_0} \log \frac{12KM}{\delta}, \quad 1 \leqslant k \leqslant K,$$

то с вероятностью не менее $(1-\delta)$ *по обучающим выборкам* $S_n \sim \mathcal{D}^{\otimes n}$ *выполнено*

$$\mathcal{E}(\widehat{f}) \leqslant \mathbb{P}_X(\widehat{f}(X) \neq f^*(X)) \lesssim \delta + \left(\frac{M^3 \log(12KM/\delta)}{n}\right)^{\frac{\alpha\beta}{\alpha\beta/p+(2\alpha+d)}}. \tag{1.12}$$

Здесь и далее в этой главе $g(n) \lesssim h(n)$ означает, что существует некоторая константа $c > 0$ такая, что $g(n) \leqslant ch(n)$ для всех натуральных значений n.

Оценки (1.11) являются оптимальными с точностью до логарифмического множителя (см. [35, Теорема 3.2] для случая отделенной от нуля плотности, [35, Теорема 4.1] для случая компактного носителя распределения, т. е. $p = 1$ в (A6), и [12, Теорема 4.5] для общего случая). Также, если плотность отделена от нуля, то можно положить $p = \infty$. Тогда неравенство (1.12) трансформируется в

$$\mathbb{P}_X(\widehat{f}(X) \neq f^*(X)) \lesssim \delta + \left(\frac{M^3 \log(12KM/\delta)}{n}\right)^{\frac{\alpha\beta}{2\alpha+d}},$$

воспроизводя результат Теоремы 7 из статьи [10].

1.6.2. Сравнение с методом k ближайших соседей

Для взвешенного метода k ближайших соседей справедлива следующая поточечная оценка.

Теорема 2. *Допустим, что выполнены условия* (A1), (A3) *и* (A5). *Зафиксируем класс* $m \in \{1, \ldots, M\}$ *и тестовую точку* $x \in \mathcal{X}$. *Тогда для оценки метода взвешенных ближайших соседей* $\widetilde{\eta}_m^w(x)$, *определенной в* (1.4) *и* (1.5), *с вероятностью хотя бы* $(1-\delta)$ *по обучающим выборкам* $S_n \sim \mathcal{D}^{\otimes n}$ *выполнено*

$$|\eta_m(x) - \widetilde{\eta}_m^w(x)| \leqslant \frac{L}{(n\varkappa p(x))^{\alpha/d}}\big(2\mathsf{k} + 4\log(2/\delta)\big)^{\alpha/d} + \sqrt{\frac{\log(4/\delta)}{\mathsf{k}}},$$

для любых k *и* $\delta \in (0,1)$, *удовлетворяющих неравенству*

$$\left(\frac{2\mathsf{k} + 4\log(1/\delta)}{n\varkappa p(x)}\right)^{\alpha/d} \leqslant r_0.$$

Результат Теоремы 2 улучшает оценку для регрессии с помощью метода ближайших соседей, полученную в [11], так как получено неравенство больших уклонений для $|\eta_m(x) - \widetilde{\eta}_m^w(x)|$ вместо верхней оценки на математическое ожидание $\mathbb{E}_{S_n}|\eta_m(x) - \widetilde{\eta}_m^w(x)|$. Для случая отделенной от нуля плотности Теорема 2 и неравенство Бонферрони позволяют немедленно получить оценку на моменты

$$\mathbb{E}_{S_n}\mathbb{E}_X \max_{1\leqslant m\leqslant M} |\eta_m(X) - \widetilde{\eta}_m^w(X)|^r \lesssim \left(\frac{\mathsf{k}\log M}{n}\right)^{\alpha r/d} + \left(\frac{\log M}{\mathsf{k}}\right)^{r/2}$$

для любого $r > 0$. Если выполнено условие малого шума, это влечет оценку на риск классификатора $\widehat{f}^{(\mathsf{k}-NN)}(x) = \operatorname*{argmax}_{1\leqslant m\leqslant M} \widetilde{\eta}_m^w(x)$:

$$\mathbb{E}_{S_n}\mathcal{E}\left(\widehat{f}^{(\mathsf{k}-NN)}(x)\right) \lesssim \left(\frac{\log M}{n}\right)^{\frac{\alpha(1+\beta)}{2\alpha+d}},$$

при условии, что $\mathsf{k} \asymp n^{2\alpha/(2\alpha+d)}$.

В случае отделенной от нуля плотности метод ближайших соседей при правильном выборе параметра достигает оптимального порядка сходимости $O(n^{-(1+\beta)/(2\alpha+d)})$, в то время как в порядке сходимости оценок Алгоритма 1 есть дополнительный логарифмический множитель.

Это легко объясняется тем фактом, что в случае $p(x) \geqslant p_0$ оптимальное значение параметра $\mathsf{k} \asymp n^{d/(2\alpha+d)}$ универсально для всех точек $x \in \mathcal{X}$. Однако в этом случае Алгоритм 1 все равно агрегирует несколько оценок, и множитель $\log n$ может быть рассмотрен как плата за адаптацию. Тем не менее, Алгоритм 1 адаптируется к неизвестному параметру гладкости $\alpha \in (0, 1]$ из условия (A3), в то время как оптимальный выбор параметра k классификатора $\widehat{f}^{(\mathsf{k}-NN)}$ основан на априорном знании α.

Ситуация радикально меняется в случае, когда плотность $p(x)$ должна удовлетворять условиям (A5) и (A6), но при этом может быть сколь угодно близка к нулю. В [12, Теоремы 4.3 и 4.5] было показано, что универсальный выбор параметра в методе ближайших соседей для всех тестовых точек ведет к субоптимальной скорости сходимости $O(n^{-\frac{\alpha(1+\beta)}{\alpha(1+\beta)/p+2\alpha+d}})$, в то время как Теорема 1 гарантирует, что Алгоритм 1 имеет минимаксно оптимальную скорость сходимости с точностью до логарифмического множителя. В [12, Теоремы 4.4 и 4.5] было показано, что если для каждой тестовой точки значение параметра $\mathsf{k} = \mathsf{k}(x)$ выбирается согласно правилу $\mathsf{k}(x) \asymp (np(x))^{2\alpha/(2\alpha+d)}$, то метод ближайших соседей достигает той же скорости сходимости, что и Алгоритм 1. Однако такой выбор значения параметра зависит от значения плотности $p(X)$ в тестовой точке X, которое неизвестно. Более того, задача оценки плотности гораздо более сложная, чем задача классификации, особенно в многомерном пространстве, поэтому получить качественную оценку $p(X)$ может быть затруднительно, тем более в случаях, когда размер обучающей выборки ограничен.

Глава 2
Восстановление многообразий

2.1. Обзор литературы

В данной главе рассмотрена задача восстановления гладкого многообразия низкой размерности по конечной выборке. Данная задача имеет большое практическое значение, а также интересна с теоретической точки зрения. Например, в задачах регрессии или частичного обучения зачастую признаки, хотя и лежат в пространстве высокой размерности, занимают лишь некоторую окрестность гладкого многообразия малой размерности. В этом случае восстановление многообразия позволяет надеяться, что качество предсказания будет зависеть от внутренней размерности, а не от внешней, тем самым преодолев проклятие размерности.

В начале 21-го века популярность задачи привела к созданию ряда новых алгоритмов снижения размерности, таких как Isomap [38], LLE (локально-линейное вложение) [39] и его модификация MLLE [40], Laplacian Eigenmaps (собственные отображения лапласиана) [41] и t-SNE (t-распределенное стохастическое вложение соседей) [42]. Недавние результаты в области оценки многообразий включают в себя интерполяцию на многообразиях с помощью геометрического мульти-разрешения [15], локально полиномиальных оценок [16] и численного решения уравнений в частных производных [43]. Отметим, что многие из алгоритмов восстановления многообразий предполагают, что наблюдения лежат на самом многообразии или в непосредственной его близости, то есть в малой окрестности многообразия $\mathcal{M}^*$, которая быстро сужается с ростом размера выборки n. К сожалению, на практике такая ситуация встречается далеко не всегда, и часто статистику предоставляется выборка из зашумленных (неточных) наблюдений. Наличие шума в данных приводит к заметному падению качества результатов. Поэтому важным направлением в задаче восстановления многообразий является проецирование неточных наблюдений на многообразие. Этот шаг можно рассматривать как предварительный для дальнейшего анализа данных на многообразии. Примеры алгоритмов проецирования данных на многообразие можно найти в работах [14, 44, 45]. Важно помнить, что в большинстве случаев такие алгоритмы действуют локально, то есть восстанавливают проекцию точки на многообразие, рассматривая ее окрестность некоторого радиуса h. Проблема данного подхода состоит в том, что h обязан быть больше амплитуды шума M, что не всегда приводит к оптимальным оценкам. Исключение составляют методы, основанные на решении задачи оптимизации, например, метод сдвига к среднему [46, 47] и его модификации [44, 48, 49]. Этот метод можно рассматривать как обобщенный EM-алгоритм, примененный к ядерной оценке плотности, основанной на выборке из неточных наблюдений (см. [50]). В силу этого метод сдвига к среднему подвержен проклятию размерности, и известные в литературе порядки сходимости этого метода зависят от внешней размерности, а не от внутренней. К нашему сведению, только в работах [17, 18] не предполагается, что амплитуда шума стремится к нулю с ростом объема выборки. Платой за это являются строгие предположения о распределении шума. Например, в работе [17] авторы считают, что шум имеет равномерное распределение в подмножестве касательного пространства. Не умаляя значимости впечатляющий результатов, полученных в [17], отметим, что данное предположение вряд ли будет выполнено на практике. Таким образом, можно заключить, что в настоящее время в литературе изучены “экстремальные”случаи, в одном из которых распределение шума принадлежит широкому классу, но амплитуда шума настолько мала, что шум может быть проигнорирован, а в другом амплитуда шума велика, но при этом распределение шума должно быть известно в точности. В данной диссертации рассматривается задача восстановления гладкого многообразия при достаточно слабых и реалистичных предположениях о распределении шума.

2.2. Восстановление гладких многообразий и структурная адаптация

Далее будет рассмотрена модель с аддитивным шумом. Предположим, что имеется простая выборка $\mathbb{Y}_n = (Y_1, \dots, Y_n) \subset \mathbb{R}^D$ многомерных наблюдений, сгенерированных из модели

$$Y = X + \varepsilon, \tag{2.1}$$

где $X \in \mathbb{R}^D$ – точка на гладком многообразии $\mathcal{M}^*$ низкой размерности $d \ll D$, $\varepsilon \in \mathbb{R}^D$ – ограниченный случайный многомерный шум с нулевым средним Задача состоит в восстановлении многообразия $\mathcal{M}^*$ и точек $X_1, \dots, X_n$. Предположения о распределении шума играют ключевую роль в теоретическом анализе. Обычно предполагается, что шум ограничен, то есть $\|\varepsilon\| \leqslant M$ почти-наверное, и имеет относительно малую амплитуду M. Если значение амплитуды меньше рича многообразия[1], то естественным является разложение вектора ε на касательную и нормальную составляющие. Влияние этих двух компонент разное, что приводит к рассмотрению модели анизотропного шума. С этой целью введем параметр b, характеризующий касательную компоненту (см. (A3)). Пара параметров (M, b) лучше характеризует структуру шума, чем просто амплитуда M и позволяет проанализировать влияние анизотропности шума на скорость сходимости оценок.

Как ранее отмечалось, многие методы проецирования неточных наблюдений на скрытое многообразие используют непараметрическое сглаживание. Если окрестность, по которой производится усреднение, имеет форму шара, это приводит к ограничению, что амплитуда шума должна быть значительно меньше радиуса окрестности (см., например, [8, 14–16]). Подобная проблема появляется и в задаче снижения эффективной размерности. Использование вытянутой в ортогональном направлении окрестности позволяет сконструировать асимптотически оптимальные оценки (см., например, [51] и [21]). В диссертации развивается идея *структурной адаптации*, предложенная в [20, 21]. Однако изложенный в диссертации материал не следует из работ [20] и [21], рассматривавших задачу обучения с учителем. Рассматриваемая в данной главе задача, задача обучения без учителя, потребовала существенных изменений в алгоритме, а также разработки полностью другого подхода к теоретическому анализу. Также отметим, что в работах [20, 21] восстанавливалось линейное подпространство, в то время как в данной главе рассматривается нелинейный случай.

2.3. Научный вклад

Кратко опишем идею алгоритма и основные результаты данной главы. Многие алгоритмы оценки гладкого многообразия (см., например, [8, 14, 19, 49]) действуют в итерационной манере, и наш метод не является исключением. Предположим, что имеются некоторые оценки $\widehat{\mathbf{\Pi}}_1^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$ проекторов на касательные к $\mathcal{M}^*$ подпространства в точках $X_1, \dots, X_n$. Несмотря на то, что эти оценки могут быть неоптимальны, они содержат некоторую информацию, которая может быть использована для построения начальных оценок $\widehat{X}_1^{(0)}, \dots, \widehat{X}_n^{(0)}$. С другой стороны, оценки $\widehat{X}_1^{(0)}, \dots, \widehat{X}_n^{(0)}$ могут быть использованы для построения новых оценок $\widehat{\mathbf{\Pi}}_1^{(1)}, \dots, \widehat{\mathbf{\Pi}}_n^{(1)}$ проекторов на касательные пространства, которые несколько лучше, чем $\widehat{\mathbf{\Pi}}_1^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$. Повторяя описанные два шага, можно последовательно улучшать оценки точек $X_1, \dots, X_n$ и многообразия $\mathcal{M}^*$. Мы будем называть такой алгоритм SAME (structure-adaptive manifold estimation или структурно-адаптивное оценивание многообразий). Далее будет доказано, что с помощью SAME можно построить оценки $\widehat{X}_1, \dots, \widehat{X}_n$ точек $X_1, \dots, X_n$ и оценку $\widehat{\mathcal{M}}$ многообразия $\mathcal{M}^*$ такие, что с большой вероятностью выполнено

$$\max_{1 \leqslant i \leqslant n} \|\widehat{X}_i - X_i\| \lesssim \frac{Mb \vee Mh \vee h^2}{\varkappa} + \sqrt{\frac{D(h^2 \vee M^2) \log n}{nh^d}}, \tag{Теорема 3}$$

[1] Определение рича многообразия дано в Главе 2.4.

$$d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \lesssim \left(\frac{M^2 b^2}{\varkappa^3} \vee \frac{h^2}{\varkappa}\right) + \sqrt{\frac{D(h^4/\varkappa^2 \vee M^2)\log n}{nh^d}}, \qquad \text{(Теорема 4)}$$

где $h \gtrsim \left((D\log n/n)^{1/d} \vee (DM^2\varkappa^2 \log n/n)^{1/(d+4)}\right)$, а M и, возможно, b убывают достаточно быстро с ростом n. Здесь h – характерный размер окрестности, который можно настраивать, $\varkappa$ – нижняя граница рича $\mathcal{M}^*$ (см. Главу 2.4). Более того, предложенный алгоритм оценивает проекторы $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$ на касательные пространства в точках $X_1, \dots, X_n$. Оценки $\widehat{\mathbf{\Pi}}_1, \dots, \widehat{\mathbf{\Pi}}_n$, получаемые с помощью этого алгоритма, с большой вероятностью удовлетворяют неравенству

$$\max_{1\leqslant i\leqslant n} \|\widehat{\mathbf{\Pi}}_i - \mathbf{\Pi}(X_i)\| \lesssim \frac{h}{\varkappa} + h^{-1}\sqrt{\frac{D(h^4/\varkappa^2 \vee M^2)\log n}{nh^d}}. \qquad \text{(Теорема 3)}$$

Здесь и далее $\|\boldsymbol{A}\|$ обозначает операторную норму матрицы $\boldsymbol{A}$. Отношение порядка $f(n) \lesssim g(n)$ означает, что существует константа $c > 0$, не зависящая от n, такая что $f(n) \leqslant cg(n)$. Через $d_H(\cdot,\cdot)$ обозначено расстояние Хаусдорфа, определяемое как

$$d_H(\mathcal{M}_1, \mathcal{M}_2) = \inf\left\{\varepsilon > 0 : \mathcal{M}_1 \subseteq \mathcal{M}_2 \oplus \mathcal{B}(0,\varepsilon),\ \mathcal{M}_2 \subseteq \mathcal{M}_1 \oplus \mathcal{B}(0,\varepsilon)\right\},$$

где $\oplus$ означает сумму Минковского и $\mathcal{B}(0,r)$ – евклидов шар в $\mathbb{R}^D$ радиуса r.

Подобрав оптимальное значение h, можно получить неравенства

$$\max_{1\leqslant i\leqslant n} \|\widehat{X}_i - X_i\| \lesssim \frac{Mb}{\varkappa} \vee \frac{1}{\varkappa}\left(\frac{D\varkappa^2\log n}{n}\right)^{\frac{2}{d+2}} \vee \frac{M}{\varkappa}\left(\frac{DM^2\varkappa^2\log n}{n}\right)^{\frac{1}{d+4}},$$

$$d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \lesssim \frac{M^2b^2}{\varkappa^3} \vee \frac{1}{\varkappa}\left(\frac{D\log n}{n}\right)^{\frac{2}{d}} \vee \frac{1}{\varkappa}\left(\frac{DM^2\varkappa^2\log n}{n}\right)^{\frac{2}{d+4}}$$

и

$$\max_{1\leqslant i\leqslant n} \|\widehat{\mathbf{\Pi}}_i - \mathbf{\Pi}(X_i)\| \lesssim \frac{1}{\varkappa}\left(\frac{D\log n}{n}\right)^{\frac{1}{d}} \vee \frac{1}{\varkappa}\left(\frac{DM^2\varkappa^2\log n}{n}\right)^{\frac{1}{d+4}}.$$

Заметим, что оптимальное значение h гораздо меньше, чем возможное значение $n^{-2/(3d+8)}$ амплитуды шума M. Помимо этого, доказана нижняя оценка на точность восстановления многообразия

$$\inf_{\widehat{\mathcal{M}}} \sup_{\mathcal{M}^*} \mathbb{E} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim \frac{M^2b^2}{\varkappa^3} \vee \varkappa^{-1}\left(\frac{M^2\varkappa^2\log n}{n}\right)^{\frac{2}{d+4}} \qquad \text{(Теорема 5)}$$

ранее не появлявшаяся в литературе. Здесь инфинум берется по всевозможным оценкам $\widehat{\mathcal{M}}$ многообразия $\mathcal{M}^*$, удовлетворяющего некоторым условиям регулярности, описанным в Теореме 5. Теорема 5, вместе с Теоремой 1 из работы [52], где авторы доказали минимаксную нижнюю оценку $\inf\limits_{\widehat{\mathcal{M}}} \sup\limits_{\mathcal{M}^*} \mathbb{E} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim (\log n/n)^{2/d}$, означает, что получаемые с помощью предложенного алгоритма оценки являются оптимальными.

2.4. Модельные предположения

Напомним, что в данной главе рассматриваем модель (2.1), где носителем распределения X является гладкое многообразие $\mathcal{M}^*$, а распределение шума ε будет описано далее в этой главе. Во-первых, необходимо потребовать регулярности многообразия $\mathcal{M}^*$. Мы предполагаем, что истинное многообразие $\mathcal{M}^*$ принадлежит классу $\mathfrak{M}^d_\varkappa$ дважды дифференцируемых,

компактных, линейно связных многообразий без края, которые содержатся в шаре $\mathcal{B}(0,R)$, имеют рич не менее $\varkappa$ и размерность d:

$$\begin{aligned}\mathcal{M}^* \in \mathfrak{M}^d_\varkappa = \big\{\mathcal{M} \subset \mathbb{R}^D : \mathcal{M} &\text{ – компактное линейно-связное} \\ &\text{многообразие без края}, \mathcal{M} \in \mathcal{C}^2, \mathcal{M} \subseteq \mathcal{B}(0,R), \\ &\operatorname{reach}(\mathcal{M}) \geqslant \varkappa, \dim(\mathcal{M}) = d < D\big\}.\end{aligned} \tag{A1}$$

Рич многообразия $\mathcal{M}$ определяется как супремум таких $r > 0$, что любая точка в $\mathcal{M} \oplus \mathcal{B}(0,R)$ имеет единственную проекцию на $\mathcal{M}$. Условие $\operatorname{reach}(\mathcal{M}^*) \geqslant \varkappa$ является условием регулярности многообразия $\mathcal{M}^*$. В частности, согласно [17, Лемма 3], кривизна любой геодезической кривой на $\mathcal{M}^*$ не превосходит $1/\varkappa$.

Во-вторых, плотность $p(x)$ вектора X (относительно d-мерной меры Хаусдорфа на $\mathcal{M}^*$) удовлетворяет неравенствам:

$$\begin{aligned}&\exists\, p_1 \geqslant p_0 > 0 : \forall x \in \mathcal{M}^* \quad p_0 \leqslant p(x) \leqslant p_1, \\ &\exists L \geqslant 0 : \forall\, x, x' \in \mathcal{M}^* \quad |p(x) - p(x')| \leqslant \frac{L\|x - x'\|}{\varkappa}.\end{aligned} \tag{A2}$$

Помимо условий на регулярность многообразия $\mathcal{M}$ и на распределение вектора X, необходимо наложить условия и на распределение шума ε. Далее предполагаем, что условное распределение $(\varepsilon|X)$ имеет следующие свойства:

$$\begin{aligned}&\mathbb{E}(\varepsilon \,\big|\, X) = 0,\ \|\varepsilon\| \leqslant M < \varkappa, \\ &\|\boldsymbol{\Pi}(X)\varepsilon\| \leqslant \frac{Mb}{\varkappa} \qquad \mathbb{P}(\cdot \,\big|\, X)\text{-почти-наверное},\end{aligned} \tag{A3}$$

где $\boldsymbol{\Pi}(X)$ – проектор на касательное пространство $\mathcal{T}_X\mathcal{M}^*$ многообразия $\mathcal{M}^*$ в точке X. Статистическая модель с $\mathcal{M}^* \in \mathfrak{M}^d_\varkappa$ и ограниченным шумом активно изучалась в литературе (см., например, [15–17, 19, 53]). В работе [54] авторы рассматривают гауссовский шум, который является неограниченным, но авторы [54] обуславливаются на событии $\max\limits_{1\leqslant i\leqslant n} \|\varepsilon_i\| \leqslant \varkappa$, что фактически сводит задачу к случаю ограниченного шума. В рассматриваемой в диссертации постановке вводится дополнительный параметр $b \in [0, \varkappa]$, характеризующий уклонение шума в касательных направлениях.

Как уже отмечалось, пара параметров (M, b) дает более полное представление о структуре шума, нежели просто амплитуда M. Случай $b = 0$ соответствуют ортогональному шуму, который рассматривался в [16, 17]. Если $b = \varkappa$, то шум просто предполагается ограниченным с амплитудой M. В данной диссертации доказываются верхние оценки на точность восстановления многообразия для всех пар (M, b), удовлетворяющих условиям:

$$\begin{cases} M \leqslant An^{-\frac{2}{3d+8}}, \\ M^3 b^2 \leqslant \alpha\varkappa \left[\left(\frac{D\log n}{n}\right)^{\frac{4}{d}} \vee \left(\frac{DM^2\varkappa^2 \log n}{n}\right)^{\frac{4}{d+4}} \right], \end{cases} \tag{A4}$$

где A и α – некоторые положительные константы. Среди всех пар (M, b), удовлетворяющих (A4), можно выделить следующие. Первая соответствует максимально возможной амплитуде:

$$\begin{aligned}&M = M(n) \leqslant An^{-\frac{2}{3d+8}}, \\ &b = b(n) \leqslant \frac{\sqrt{\alpha\varkappa}}{A^{3/2}} \left[\left(\frac{D\log n}{n}\right)^{\frac{1}{d}} \vee \left(\frac{DM^2\varkappa^2 \log n}{n}\right)^{\frac{1}{d+4}} \right].\end{aligned} \tag{A4.1}$$

Вторая – максимально возможному значению b:

$$b = \varkappa, \quad M = M(n) \leqslant \left(\frac{D^4 \alpha^{d+4}}{\varkappa^{d-4}} \right)^{\frac{1}{3d+4}} n^{-\frac{4}{3d+4}}. \tag{A4.2}$$

Если выполнено условие (A4.1), то шум *почти* ортогонален многообразию. В этом случае X оказывается близок к проекции $\pi_{\mathcal{M}^*}(Y)$ наблюдения Y на $\mathcal{M}^*$. Здесь и далее для любого замкнутого множества $\mathcal{M}$ $\pi_{\mathcal{M}}(x)$ обозначает проекцию x на $\mathcal{M}$. Таким образом, оценки точек $X_1, \dots, X_n$ могут также рассматриваться как оценки проекций $Y_1, \dots, Y_n$ на $\mathcal{M}^*$. Также допускается, что M может убывать как $n^{-2/(3d+8)}$. Это условие обсуждается более подробно после Теоремы 3. К нашему сведению, только в [17, 18] авторы накладывают более слабые условия на амплитуду шума. На первый взгляд, условие (A4.1) очень похоже на условие ортогонального шума $b = 0$. Однако теоретический анализ показывает, что существующие нижние оценки на точность восстановления многообразия в случае перпендикулярного шума отличаются от доказанной в этой диссертации оценки для случая шума, удовлетворяющего (A4.1). Подробное сравнение нижних оценок приведено в Главе 2.6.

Наконец, если выполнено условие (A4.2), специальной анизотропной структуры шума не предполагается. Однако в этом случае необходимо наложить более строгое условие на амплитуду шума. Тем не менее, далее будет доказано, что в случае ограниченного шума результат [19, Теорема 2.7] может быть улучшен, если дополнительно предположить, что $\log p(x)$ Липшицев. Подробное сравнение приведено в Главе 2.6.

2.5. Восстановление гладкого многообразия методом адаптивных весов

В данной главе описана новая процедура для оценивания многообразий на основе стандартной для непараметрической статистики техники сглаживания и идеи структурной адаптации. Один из самых известных методов в непараметрической статистике – это взвешенное среднее

$$\widehat{X}_i^{(loc)} = \frac{\sum\limits_{j=1}^{n} w_{ij}^{(loc)} Y_j}{\sum\limits_{j=1}^{n} w_{ij}^{(loc)}}, \quad 1 \leqslant i \leqslant n, \tag{2.2}$$

где локализующие веса $w_{ij}^{(loc)}$ вычисляются по формуле

$$w_{ij}^{(loc)} = \mathcal{K}\left(\frac{\|Y_i - Y_j\|^2}{h^2} \right), \quad 1 \leqslant i, j \leqslant n,$$

$\mathcal{K}(\cdot)$ – сглаживающее ядро, $h = h(n)$ – параметр, который необходимо настроить. В данной главе рассматривается ядро $\mathcal{K}(t) = e^{-t}$.

Замечание 1. *Вместо $\mathcal{K}(t) = e^{-t}$ можно взять любую дважды дифференцируемую, монотонно убывающую на $\mathbb{R}_+$ функцию, первая и вторая производные которых либо экспоненциально убывают, либо имеют компактный носитель. Выбор $\mathcal{K}(t) = e^{-t}$ сделан во избежание дальнейшего усложнения доказательств.*

Оценка (2.2) имеет очевидное ограничение. Рассмотрим пару индексов (i, j), такую что $\|X_i - X_j\| \leqslant h$, и выберем $h = h(n)$ порядка $(\log n/n)^{1/d}$, которое является оптимальным при малом шуме (см. [19, Предложение 5.1] и [16, Теорема 6]). Но если амплитуда шума M становится больше $(\log n/n)^{1/d}$, то $M > h$, и веса $w_{ij}^{(loc)}$ дают неправильную информацию об

окрестности точки X_i, то есть вес $w_{ij}^{(loc)}$ может быть близким к нулю даже в случае, когда $\|X_i - X_j\| < h$. Из-за этого оценка (2.2) имеет большую дисперсию при h порядка $(\log n/n)^{1/d}$, поэтому необходимо увеличить h, что неизбежно ведет к увеличению смещения оценки.

Таким образом, веса $w_{ij}^{(loc)}$ непригодны для построения качественных оценок и должны быть модифицированы. Зафиксируем произвольное i от 1 до n. "Идеальные" веса w_{ij} должны обладать тем свойством, что они близки к нулю в том и только том случае, когда $\|X_i - X_j\| > h$. К сожалению, подсчитать норму $\|X_i - X_j\|$ не представляется возможным, так как наблюдения $X_1, \dots, X_n$ скрыты. Однако предположим, что проектор $\mathbf{\Pi}(X_i)$ на касательное пространство $\mathcal{T}_{X_i}\mathcal{M}^*$ известен. Тогда можно рассмотреть веса вида

$$w_{ij}(\mathbf{\Pi}(X_i)) = \mathcal{K}\left(\frac{\|\mathbf{\Pi}(X_i)(Y_i - Y_j)\|^2}{h^2}\right), \quad 1 \leqslant j \leqslant n,$$

чтобы уменьшить влияние ортогональной компоненты шума. Норма $\|\mathbf{\Pi}(X_i)(Y_i - Y_j)\|$ оказывается ближе к $\|X_i - X_j\|$, чем $\|Y_i - Y_j\|$, особенно если размерность D велика. Таким образом, вместо шара $\{Y : \|Y - Y_i\| \leqslant h\}$ мы рассматриваем цилиндр $\{Y : \Pi_i(Y_i - Y_j)\}$, где $\mathbf{\Pi}_i$ – проектор, близкий к $\Pi(X_i)$ по операторной норме. Необходимо лишь убедиться в том, что цилиндр не пересекается с $\mathcal{M}^*$ несколько раз. С этой целью мы рассматриваем веса

$$w_{ij}(\mathbf{\Pi}_i) = \mathcal{K}\left(\frac{\|\mathbf{\Pi}_i(Y_i - Y_j)\|^2}{h^2}\right) \mathbb{1}\left(\|Y_i - Y_j\| \leqslant \tau\right), \quad 1 \leqslant j \leqslant n, \tag{2.3}$$

где $\tau < \varkappa$.

Модифицированные веса (2.3) требуют качественной оценки $\mathbf{\Pi}_i$ проектора $\mathbf{\Pi}(X_i)$. Вопрос в том, как получить качественную оценку $\mathbf{\Pi}_i$. Мы используем следующую стратегию. Пусть даны начальные оценки $\widehat{\mathbf{\Pi}}_1{}^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$ проекторов $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$ (которые, на самом деле, могут быть довольно неточными). Возьмем достаточно большое h_0 и вычислим взвешенные средние $\widehat{X}_1^{(1)}, \dots, \widehat{X}_n^{(1)}$ с весами (2.3) и параметром h_0. Полученные оценки могут быть использованы при построении оценок $\widehat{\mathbf{\Pi}}_1^{(1)}, \dots, \widehat{\mathbf{\Pi}}_n^{(1)}$ проекторов $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$, которые оказываются лучше, чем $\widehat{\mathbf{\Pi}}_1{}^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$. После этого описанные два шага повторяются с параметром $h_1 < h_0$. Таким образом, получаем итеративную процедуру, описанную подробно в Алгоритме 2.

Вычислительная сложность Алгоритма 2 составляет $O(n^2D^2K + nD^3K)$, включая $O(n^2D^2)$ операций, необходимых для обновления весов $w_{ij}^{(k)}$, $1 \leqslant i, j \leqslant n$, и оценок $\widehat{X}_i^{(k)}$ и $\widehat{\mathbf{\Sigma}}_i^{(k)}$, $1 \leqslant i \leqslant n$, на каждой итерации, а также $O(nD^3)$ операций, чтобы обновить оценки проекторов $\widehat{\mathbf{\Pi}}_i^{(k)}$, $1 \leqslant i \leqslant n$, на каждой итерации.

2.6. Теоретические свойства алгоритма SAME

В данной главе сформулированы основные теоретические свойства Алгоритма 2. Здесь и далее $\|\boldsymbol{A}\|$ обозначает операторную норму матрицы $\boldsymbol{A}$. Отношение эквивалентности $f(n) \asymp g(n)$ означает $f(n) \lesssim g(n) \lesssim f(n)$.

Теорема 3. *Пусть выполнены условия* (A1), (A2), (A3), *and* (A4) . *Пусть начальные приближения* $\widehat{\mathbf{\Pi}}_1^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$ *проекторов* $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$ *таковы, что с вероятностью не менее* $1 - n^{-1}$ *выполнено*

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(0)} - \mathbf{\Pi}(X_i)\| \leqslant \frac{\Delta h_0}{\varkappa}$$

с константой Δ, *такой что* $\Delta h_0 \leqslant \varkappa/4$, *и* $h_0 = C_0/\log n$, *где* $C_0 > 0$ – *некоторая константа. Выберем* $\tau = 2C_0/\sqrt{\log n}$ *и возьмем* $a \in (1, 2]$. *Если объем выборки* n *превосходит константу* N_Δ, *зависящую от* Δ, *и* $h_K \gtrsim \left((D\log n/n)^{1/d} \vee (DM^2\varkappa^2 \log n/n)^{1/(d+4)}\right)$ *(с достаточно большой скрытой константой, превосходящей* 1*), существует такое* γ, *что после*

Алгоритм 2 Структурно-адаптивная оценка многообразий (SAME)

1: Выборка $\mathbb{Y}_n = (Y_1, \dots, Y_n)$, начальные приближения $\widehat{\mathbf{\Pi}}_1^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$ проекторов $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$, число итераций $K+1$, константы h_0, τ, $a > 1$ и $\gamma > 0$ заданы.

2: **Цикл** k от 0 до K **:**

3: Вычислить веса $w_{ij}^{(k)}$ по формуле

$$w_{ij}^{(k)} = \mathcal{K}\left(\frac{\|\widehat{\mathbf{\Pi}}_i^{(k)}(Y_i - Y_j)\|^2}{h_k^2}\right) \mathbb{1}\left(\|Y_i - Y_j\| \leqslant \tau\right), \quad 1 \leqslant i, j \leqslant n\,.$$

4: Вычислить оценки

$$\widehat{X}_i^{(k)} = \sum_{j=1}^n w_{ij}^{(k)} Y_j \Big/ \left(\sum_{j=1}^n w_{ij}^{(k)}\right), \quad 1 \leqslant i \leqslant n\,. \tag{2.4}$$

5: Если $k < K$, для каждого i от 1 до n, определить множество $\mathcal{J}_i^{(k)} = \{j : \|\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)}\| \leqslant \gamma h_k\}$ и вычислить

$$\widehat{\mathbf{\Sigma}}_i^{(k)} = \sum_{j \in \mathcal{J}_i^{(k)}} (\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)})(\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)})^T, \quad 1 \leqslant i \leqslant n.$$

6: Если $k < K$, для каждого i от 1 до n положить $\widehat{\mathbf{\Pi}}_i^{(k+1)}$ равным проектору на линейную оболочку собственных векторов $\widehat{\mathbf{\Sigma}}_i^{(k)}$, отвечающих d наибольшим собственным значениям.

7: Если $k < K$, положить $h_{k+1} = a^{-1} h_k$.

8: **Конец цикла**

9: **Вернуть** оценки $\widehat{X}_1 = \widehat{X}_1^{(K)}, \dots, \widehat{X}_n = \widehat{X}_n^{(K)}$.

K итераций Алгоритм 2 выдает оценки $\widehat{X}_1, \dots, \widehat{X}_n$, такие что с вероятностью хотя бы $1 - (5K+4)/n$ выполнено

$$\max_{1 \leqslant i \leqslant n} \|\widehat{X}_i - X_i\| \lesssim \frac{Mb \vee Mh_K \vee (1 + \Phi_{M,b,h_K,\varkappa}) h_K^2}{\varkappa} + \sqrt{\frac{D(h_K^2 \vee M^2)\log n}{n h_K^d}},$$

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(K)} - \mathbf{\Pi}(X_i)\| \lesssim \Psi_{M,b,h_K,\varkappa} \left(\frac{h_K}{\varkappa} + h_K^{-1}\sqrt{\frac{D(h_K^4/\varkappa^2 \vee M^2)\log n}{n h_K^d}}\right),$$

где

$$\begin{aligned}
\Phi_{M,b,h_K,\varkappa} &= \frac{M^3(1 + b/h_K)^2}{h_K^2 \varkappa} + \frac{M^2(1 + b/h_K + \sqrt{\log h_K^{-1}})}{\varkappa h_K} + \frac{M h_K^2}{\varkappa^3} \\
&\lesssim \alpha + o(1), \quad n \to \infty, \\
\Psi_{M,b,h_K,\varkappa} &= \left(1 + \frac{M(1 + b/h_K) \vee (1 + \Phi_{M,b,h_K,\varkappa}) h_K}{\varkappa}\right)^{d+1} (1 + \Phi_{M,b,h_K,\varkappa}) \\
&\leqslant (1+\alpha)\left(4^{d+1} + (2\sqrt{\alpha})^{d+1}\right).
\end{aligned} \tag{2.5}$$

В частности, если выбрать параметр a и число итераций K таким образом, что $h_K \asymp \left((D\varkappa^2 \log n/n)^{1/(d+2)} \vee (DM^2\varkappa^2 \log n/n)^{1/(d+4)}\right)$, то

$$\max_{1 \leqslant i \leqslant n} \|\widehat{X}_i - X_i\| \lesssim \frac{Mb}{\varkappa} + \frac{1}{\varkappa}\left(\frac{D\varkappa^2 \log n}{n}\right)^{\frac{2}{d+2}} \vee \frac{M}{\varkappa}\left(\frac{DM^2\varkappa^2 \log n}{n}\right)^{\frac{1}{d+4}}.$$

Если $h_K \asymp \big((D\log n/n)^{1/d} \vee (DM^2\varkappa^2\log n/n)^{1/(d+4)}\big)$, *то*

$$\max_{1\leqslant i\leqslant n}\|\widehat{\mathbf{\Pi}}_i^{(K)} - \mathbf{\Pi}(X_i)\| \lesssim \frac{1}{\varkappa}\left(\frac{D\log n}{n}\right)^{\frac{1}{d}} \vee \frac{1}{\varkappa}\left(\frac{DM^2\varkappa^2\log n}{n}\right)^{\frac{1}{d+4}}.$$

Заметим, что необходимое число итераций K имеет порядок $O(\log n)$, поскольку последовательность $h_1,\dots,h_K$ убывает экспоненциально. В Теореме 3, допускается, что начальные оценки проекторов $\widehat{\mathbf{\Pi}}_1^{(0)},\dots,\widehat{\mathbf{\Pi}}_n^{(0)}$ могут зависеть от $Y_1,\dots,Y_n$. Возникает естественный вопрос, как построить начальные оценки проекторов $\mathbf{\Pi}(X_1),\dots,\mathbf{\Pi}(X_n)$, удовлетворяющие условию теоремы. Для этого можно использовать результат работы [19, Предложение 5.1]. Для каждого i от 1 до n обозначим

$$\widehat{\mathbf{\Sigma}}_i^{(0)} = \frac{1}{n-1}\sum_{j\neq i}(Y_j - \overline{Y}_i)(Y_j - \overline{Y}_i)^T\mathbb{1}(Y_j \in \mathcal{B}(Y_i, h_0)),$$

где $\overline{Y}_i = \frac{1}{N_i}\sum_{j\neq i} Y_j\mathbb{1}(Y_j \in \mathcal{B}(Y_i,h_0))$, $N_i = |\{j : Y_j \in \mathcal{B}(Y_i,h)\}|$. Пусть $\widehat{\mathbf{\Pi}}_i^{(0)}$ – проектор на линейную оболочку собственных векторов матрицы $\widehat{\mathbf{\Sigma}}_i^{(0)}$, отвечающих d наибольшим собственным значениям $\widehat{\mathbf{\Sigma}}_i^{(0)}$. Следуя доказательству [19, Предложение 5.1], можно показать, что с вероятностью не менее $1-n^{-1}$ выполнено неравенство

$$\max_{1\leqslant i\leqslant n}\|\widehat{\mathbf{\Pi}}_i^{(0)} - \mathbf{\Pi}(X_i)\| \lesssim \frac{h_0}{\varkappa} + \frac{M}{h_0},$$

при условии, что $h_0 \gtrsim (\log n/n)^{1/d}$, $h_0 = h_0(n) = o(1)$ при $n\to\infty$, и n достаточно велико.

Из условия (A4) и выбора h_K в Теореме 3 следует, что $M = M(n)$ может быть порядка $h_K^{2/3} = h_K^{2/3}(n)$. Таким образом, допускается ситуация, при которой амплитуда шума M оказывается значительно больше параметра сглаживания h_K. Например, в статье [16] авторы используют локально полиномиальные оценки и требуют $M = O(h^2)$ и $h = h(n) \asymp n^{-1/d}$. В работе [19] авторы предполагают $M \lesssim \lambda(\log n/n)^{1/d}$. В статье [54] авторы рассматривают гауссовский шум $\mathcal{N}(0,\sigma^2\boldsymbol{I}_D)$ и восстанавливают многообразие $\mathcal{M}^*$ с точностью $O(\sigma\sqrt{D})$, используя выборку объема $O(\sigma^{-d})$. Это означает, что $\sigma = O(n^{-1/d})$ и, следовательно, с большой вероятностью выполнено неравенство

$$\max_{1\leqslant i\leqslant n}\|\varepsilon_i\| \lesssim n^{-1/d}\sqrt{D\log n}.$$

Похожая ситуация наблюдается в работе [49], где также рассмотрен гауссовский шум $\mathbb{N}(0,\sigma^2\boldsymbol{I}_D)$. Используя ядерную оценку плотности с параметром h, авторы получают верхнюю оценку

$$O\left(\sigma^2\log\sigma^{-1} + h^2 + \sqrt{\frac{\log n}{nh^D}}\right)$$

на расстояние Хаусдорфа между $\mathcal{M}^*$ и его оценкой. Чтобы первое слагаемое не было доминирующим, должно быть выполнено условие $\sigma = O(h/\sqrt{\log h^{-1}})$, которое означает, что

$$\max_{1\leqslant i\leqslant n}\|\varepsilon_i\| \lesssim h\sqrt{\frac{D\log n}{\log h^{-1}}},$$

в то время как в данной диссертации $\max_{1\leqslant i\leqslant n}\|\varepsilon_i\|$ может быть порядка $h_K^{2/3}$. Наконец, в работе [14] авторы требуют выполнения условия $M = O(h)$. Как видно из многочисленных примеров, условие (A4) достаточно мягкое.

Согласно Теореме 3, несмотря на относительно большой шум, предложенный метод успешно восстанавливает проекции точек $Y_1,\dots,Y_n$ на многообразие $\mathcal{M}^*$. Однако точность

оценки проекций несколько хуже, чем точность восстановления многообразия в терминах метрики Хаусдорфа (см. Теорему 4 ниже). Этот факт легко объяснить тем, что оценка $\widehat{X}_i$ существенно сдвинута по отношению к X_i в касательном направлении тогда как ортогональная компонента вектора $(\widehat{X}_i - X_i)$ мала. Похожий феномен наблюдался в задаче эффективного снижения размерности. Например, в статьях [20], [21] был получен порядок $n^{-2/3}$ для компоненты смещения оценки, ортогональной линейному пространству, несущему всю необходимую информацию о признаках, в то время как для смещения оценки удалось получить лишь порядок $n^{-1/2}$. Более того, слагаемое Mh_K в Теореме 3 появляется из-за корреляции между весами $w_{ij}^{(k)}$ и элементами выборки Y_j. Чтобы избавиться от этого слагаемого, необходимо рассмотреть другое ядро, которое позволит уменьшить корреляцию.

Сформулируем теперь результат о точности восстановления многообразия $\mathcal{M}^*$.

Теорема 4. *Пусть выполнены условия Теоремы 3. Рассмотрим кусочно линейную оценку*

$$\widehat{\mathcal{M}} = \left\{ \widehat{X}_i + h_K \widehat{\mathbf{\Pi}}_i^{(K)} u : 1 \leqslant i \leqslant n,\, u \in \mathcal{B}(0,1) \subset \mathbb{R}^D \right\},$$

где $\widehat{\mathbf{\Pi}}_i^{(K)}$ – проектор на d-мерное линейное пространство, полученный на K-ой итерации Алгоритма 2. Тогда, если $h_K \gtrsim \left((D\log n/n)^{1/d} \vee (DM^2\varkappa^2 \log n/n)^{1/(d+4)}\right)$ (с достаточно большой скрытой константой, превосходящей 1), с вероятностью хотя бы $1-(5K+5)/n$ выполнено

$$d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \lesssim \left(\frac{(1+\Phi_{M,b,h_K,\varkappa} + \Psi_{M,b,h_K,\varkappa}) h_K^2}{\varkappa} \vee \frac{M^2 b^2}{\varkappa^3} \right) + \sqrt{\frac{D(h_K^4/\varkappa^2 \vee M^2)\log h_K^{-1}}{n h_K^d}},$$

где $\Phi_{M,b,h_K,\varkappa}$ и $\Psi_{M,b,h_K,\varkappa}$ определены в (2.5). В частности, если a и K выбраны так, что $h_K \asymp \left((D\log n/n)^{1/d} \vee (DM^2\varkappa^2/n\log n)^{1/(d+4)}\right)$, то

$$d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \lesssim \frac{M^2 b^2}{\varkappa^3} \vee \varkappa^{-1} \left(\frac{D\log n}{n}\right)^{\frac{2}{d}} \vee \varkappa^{-1} \left(\frac{DM^2\varkappa^2 \log n}{n}\right)^{\frac{2}{d+4}}.$$

Конкретизируем результат Теоремы 4 для случаев, когда выполнены условия (A4.1) и (A4.2). Начнем с условия (A4.2). Модель с ограниченным шумом была рассмотрена в работе [19], где авторы предположили, что $\mathcal{M}^*$ удовлетворяет условию (A1), а для плотности X выполнено неравенство

$$0 < p_0 \leqslant p(x) \leqslant p_1, \quad \forall\, x \in \mathcal{M}^*,$$

при некоторых константах p_0, p_1. Заметим, что данное условие мягче, чем (A2), поскольку в (A2) дополнительно предполагается Липшицевость логарифма плотности. При сформулированных предположениях в [19, Теорема 2.7] была доказана следующая верхняя оценка для касательного комплекса Делонэ (tangential Delaunay complex, TDC):

$$d_H(\widehat{\mathcal{M}}_{TDC}, \mathcal{M}^*) \lesssim \left(\frac{\log n}{n}\right)^{2/d} + M^2 \left(\frac{\log n}{n}\right)^{-2/d},$$

при условии $M \lesssim (\log n/n)^{1/d}$. Насколько нам известно, модель, в которой выполнены условия (A1), (A2), (A3) и (A4.2) не была изучена в литературе. Заметим, что и SAME, и TDC достигают скорости сходимости $O\left(\log n/n\right)^{2/d}$ в случае шума с малой амплитудой $M \lesssim (\log n/n)^{2/d}$. Однако, если $(\log n/n)^{2/d} \lesssim M \lesssim n^{-4/(3d+4)}$, то скорость сходимости SAME (при условии, что $p(x)$ удовлетворяет (A2)) оказывается лучше, чем у TDC в случае отделенной от нуля ограниченной плотности $p(x)$.

Перейдем к рассмотрению случая почти ортогонального шума, то есть когда выполнено условие (A4.1). Такая модель ранее не рассматривалась в других работах. Наиболее близкой является модель с ортогональным шумом, изученная в [16, 17], поэтому имеет смысл сравнить ее с рассматриваемой в диссертации моделью. В работе [17] был получен порядок сходимости $O(\log n/n)^{2/(d+2)}$ в предположении, что условное распределение $(\varepsilon|X)$ равномерное на множестве $\mathcal{B}(X, M) \cap (\mathcal{T}_X\mathcal{M}^*)^\perp$. В отличие от данной диссертации, авторы не предполагали, что M стремится к нулю при $n \to \infty$, однако они накладывают гораздо более жесткое условие на распределение шума. В статье [16, Теорема 6] авторы доказали, что для локально полиномиальной оценки $\widehat{\mathcal{M}}_{LP}$ выполнено неравенство

$$d_H(\widehat{\mathcal{M}}_{LP}, \mathcal{M}^*) \lesssim \left(\frac{\log n}{n}\right)^{k/d} \vee M$$

при условии, что $\mathcal{M}^*$ – компактное многообразие без края в $\mathbb{R}^D$, $\mathcal{M}^* \in \mathcal{C}^k$, размерность $\mathcal{M}^*$ равна d, его рич не менее $\varkappa$. В случае $k = 2$ эта оценка оптимальна при малом шуме ($M \lesssim (\log n/n)^{2/d}$), но может быть улучшена, если M превышает $(\log n/n)^{2/d}$.

Результат Теоремы 4 неулучшаем в случае аддитивного шума, удовлетворяющего предположению (A3) с $b \gtrsim \left((\log n/n)^{1/d} \vee (M^2\varkappa^2 \log n/n)^{1/(d+4)}\right)$. Данное утверждение подтверждается следующей теоремой.

Теорема 5. *Пусть выборка $\mathbb{Y}_n = \{Y_1, \dots, Y_n\}$ сгенерирована из модели (2.1), где $\mathcal{M}^* \in \mathfrak{M}^d_\varkappa$, плотность $p(x)$ случайного вектора X удовлетворяет (A2) (с достаточно большими p_1, L и малым p_0), а для шума ε выполнено условие (A3). Тогда для любой оценки $\widehat{\mathcal{M}}$, выполнено неравенство*

$$\sup_{\mathcal{M}^* \in \mathfrak{M}^d_\varkappa} \mathbb{E}_{\mathcal{M}^*} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim \frac{M^2 b^2}{\varkappa^3}. \tag{2.6}$$

Более того, если дополнительно объем выборки n достаточно велик, $M\varkappa \gtrsim (\log n/n)^{2/d}$, а параметр b в (A3) удовлетворяет неравенству

$$b \gtrsim \left((\log n/n)^{1/d} \vee (M^2\varkappa^2 \log n/n)^{1/(d+4)}\right),$$

с достаточно большой скрытой константой, то для любой оценки $\widehat{\mathcal{M}}$ выполнено

$$\sup_{\mathcal{M}^* \in \mathfrak{M}^d_\varkappa} \mathbb{E}_{\mathcal{M}^*} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim \varkappa^{-1} \left(\frac{M^2\varkappa^2 \log n}{n}\right)^{\frac{2}{d+4}}. \tag{2.7}$$

В Теореме 5 предполагается, что $M \gtrsim (\log n/n)^{2/d}$. Комплементарный случай рассмотрен в работе [52], где авторы доказали нижнюю оценку

$$\inf_{\widehat{\mathcal{M}}} \sup_{\mathcal{M}^* \in \mathfrak{M}^d_\varkappa} \mathbb{E}_{\mathcal{M}^*} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim \left(\frac{\log n}{n}\right)^{2/d}$$

для бесшумного случая, которая также справедлива и для $M \lesssim (\log n/n)^{2/d}$. Из Теоремы 5 и [52, Теорема 1] следует, что алгоритм SAME позволяет построить оптимальные оценки в модели с почти ортогональным шумом. Нижние оценки (2.6) и (2.7) ранее не встречались в литературе и значительно отличаются от полученных в [17] и [16] оценок для случая ортогонального шума, то есть удовлетворяющего условию (A3) с $b = 0$.

Заключение

1. Предложен адаптивный алгоритм многоклассовой классификации, основанный на агрегации оценок метода k ближайших соседей. Алгоритм автоматически выбирает близкое к оптимальному значение k для каждой точки и каждого класса, а также адаптируется к неизвестной гладкости целевой функции.

2. Доказана верхняя оценка на избыточный риск классификатора, получаемого в результате работы алгоритма, при мягких предположениях. Это первый теоретический результат, в котором скорость сходимости совпадает с минимаксной нижней оценкой с точностью до логарифмического множителя при данных предположениях.

3. Разработана новая структурно адаптивная оценка гладкого многообразия по неточным наблюдениям. Предложенная процедура оказывается более устойчивой к ортогональному шуму, чем существующие методы.

4. Проведен теоретический анализ предложенной процедуры. Доказаны новые верхняя и нижняя оценка на точность восстановления гладкого многообразия, совпадающие с точностью до мультипликативной константы. Таким образом, подтверждается оптимальность предложенного алгоритма в минимаксном смысле.

Список литературы

1. C. H.Q. Ding and I. Dubchak. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4):349–58, 2001.
2. H. Ahn and K.-J. Kim. Corporate credit rating using multiclass classification models with order information. *World Academy of Science, Engineering and Technology*, 60:95–100, 12 2011.
3. D. Belomestny and V. Spokoiny. Spatial aggregation of local likelihood estimates with applications to classification. *The Annals of Statistics*, 35(5):2287–2311, 2007.
4. A. Ganapathiraju, J.E. Hamaker, and J. Picone. Applications of support vector machines to speech recognition. *IEEE Transactions on Signal Processing*, 52(8):2348–2355, 2004.
5. J. Kittler, R. Ghaderi, T. Windeatt, and J. Matas. Face verification via error correcting output codes. *Image and Vision Computing*, 21:1163–1169, 12 2003.
6. D. Li and D. B. Dunson. Classification via local manifold approximation. *Biometrika*, 107(4):1013–1020, 2020.
7. G. Peyré. Manifold models for signals and images. *Computer Vision and Image Understanding*, 113(2):249–260, 2009.
8. S. Osher, Z. Shi, and W. Zhu. Low dimensional manifold model for image processing. *SIAM Journal on Imaging Sciences*, 10(4):1669–1690, 2017.
9. R. J. Samworth. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012.
10. K. Chaudhuri and S. Dasgupta. Rates of convergence for nearest neighbor classification. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3437–3445. MIT Press, 2014.
11. M. Döring, L. Györfi, and H. Walk. Rate of convergence of k-nearest-neighbor classification rule. *Journal of Machine Learning Research*, 18(227):1–16, 2018.
12. S. Gadat, T. Klein, and C. Marteau. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *The Annals of Statistics*, 44(3):982–1009, 2016.
13. T. I. Cannings, T. B. Berrett, and R. J. Samworth. Local nearest neighbour classification with applications to semi-supervised learning. *The Annals of Statistics*, 48(3):1789–1814, 2020.
14. M. Hein and M. Maier. Manifold denoising. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.
15. M. Maggioni, S. Minsker, and N. Strawn. Multiscale dictionary learning: non-asymptotic bounds and robustness. *Journal of Machine Learning Research*, 17(2):1–51, 2016.
16. E. Aamari and C. Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177–204, 2019.
17. C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Minimax manifold estimation. *Journal of Machine Learning Research*, 13(43):1263–1291, 2012.
18. C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics*, 40(2):941–963, 2012.
19. E. Aamari and C. Levrard. Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Discrete & Computational Geometry*, 59(4):923–971, 2018.
20. M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *The Annals of Statistics*, 29(3):595–623, 2001.
21. M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566, 2001.
22. N. Puchkin and V. Spokoiny. An adaptive multiclass nearest neighbor classifier. *ESAIM: Probability and Statistics*, 24:69–99, 2020.
23. N. Puchkin and V. Spokoiny. Structure-adaptive manifold estimation. *Journal of Machine Learning Research*, 23(40):1–62, 2022.

24. R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 18(1):104–117, 2003.
25. A. B. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pages 303–313. Springer Berlin Heidelberg, 2003.
26. G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007.
27. P. Rigollet and A. B. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 27(4):558–575, 2012.
28. A. B. Yuditskiĭ, A. V. Nazin, A. B. Tsybakov, and N. Vayatis. Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii*, 41(4):78–96, 2005.
29. A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.
30. G. Lecué. Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli*, 19(5B):2153–2166, 2013.
31. D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, 40(3):1878–1905, 2012.
32. G. Lecué and P. Rigollet. Optimal learning with Q-aggregation. *The Annals of Statistics*, 42(1):211–224, 2014.
33. C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss. *Electronic Journal of Statistics*, 11(1):2258–2294, 2017.
34. P. Rigollet. Kullback-Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40(2):639–665, 2012.
35. J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.
36. V. Dinh, L. S. T. Ho, N. V. Cuong, D. Nguyen, and B. T. Nguyen. Learning from non-iid data: fast rates for the one-vs-all multiclass plug-in classifiers. In *Theory and applications of models of computation*, volume 9076 of *Lecture Notes in Comput. Sci.*, pages 375–387. Springer, Cham, 2015.
37. E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.
38. J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
39. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
40. Z. Zhang and J. Wang. MLLE: Modified locally linear embedding using multiple weights. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.
41. M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
42. L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
43. Z. Shi and J. Sun. Convergence of the point integral method for Laplace-Beltrami equation on point cloud. *Research in the Mathematical Sciences*, 4(1):22, 2017.
44. W. Wang and M. Á. Carreira-Perpiñán. Manifold blurring mean shift algorithms for manifold denoising. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1759–1766, 2010.
45. D. Gong, F. Sha, and G. Medioni. Locally linear denoising on image manifolds. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 265–272. PMLR, 2010.

46. K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.
47. Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.
48. U. Ozertem and D. Erdogmus. Locally defined principal curves and surfaces. *Journal of Machine Learning Research*, 12(34):1249–1286, 2011.
49. C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511–1545, 2014.
50. M. Á. Carreira-Perpiñán. Gaussian mean-shift is an em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776, 2007.
51. Y. Xia, H. Tong, W. K. Li, and L.-X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3):363–410, 2002.
52. A. K. H. Kim and H. H. Zhou. Tight minimax rates for manifold estimation under Hausdorff loss. *Electronic Journal of Statistics*, 9(1):1562–1582, 2015.
53. N. G. Trillos, D. Sanz-Alonso, and R. Yang. Local regularization of noisy point clouds: Improved global geometric estimates and data analysis. *Journal of Machine Learning Research*, 20(136):1–37, 2019.
54. C. Fefferman, S. Ivanov, Y. Kurylev, M. Lassas, and H. Narayanan. Fitting a putative manifold to noisy data. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 688–720. PMLR, 2018.