На правах рукописи

Бахтеев Олег Юрьевич

Байесовский выбор субоптимальной структуры модели глубокого обучения

05.13.17 — Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2019

Работа выполнена на кафедре интеллектуальных систем федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский институт)».

Научный руководитель:

TODO Защита состоится «25» декабря 2019 года в 13:00 на заседании диссертационного совета Д 002.073.05 при Федеральном исследовательском центре «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН) по адресу: 119333, г. Москва, ул. Вавилова, д. 40.

С диссертацией можно ознакомиться в библиотеке Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» и на сайте http://www.frccsc.ru/

Автореферат разослан TODO «__» _____ 2019 года.

Ученый секретарь
диссертационного совета Д 002.073.05
д.ф.-м.н., профессор В.В.Рязанов

Общая характеристика работы

Актуальность темы. TODO: формат библиографии? В работе рассматривается задача автоматического построения моделей глубокого обучения оптимальной и субоптимальной сложности.

Под сложностью модели понимается *минимальная длина описания* [?], т.е. минимальное количество информации, которое требуется для передачи информации о модели и о выборке. Вычисление минимальной длины описания модели является вычислительно сложной процедурой. В работе предлагается получение ее приближенной оценки, основанной на связи минимальной длины описания и *обоснованности модели* [?]. Для получения оценки обоснованности используются вариационные методы получения оценки обоснованности [?], основанные на аппроксимации неизвестного апостериорного распределения другим заданным распределением. Под субоптимальной сложностью понимается вариационная оценка обоснованности модели.

Одна из проблем построения моделей глубокого обучения — большое количество параметров моделей [?, ?]. Поэтому задача выбора моделей глубокого обучения включает в себя выбор стратегии построения модели, эффективной по вычислительным ресурсам. В работе [?] приводятся теоретические оценки построения нейросетей с использованием жадных стратегий, при которых построение модели производится итеративно последовательным увеличением числа нейронов в сети. В работе [?] предлагается жадная стратегия выбора модели нейросети с использованием релевантных распределений, т.е. параметрических распределений, оптимизация параметров которых позволяет удалить часть параметров из модели. Данный метод был также применялся в задаче построения модели метода релевантных векторов [?]. Альтернативой данным алгоритмам построения моделей являются методы, основанные на прореживании сетей глубокого обучения [?, ?, ?], т.е. на последовательном удалении параметров, не дающих существенного прироста качества модели. В работах [?, ?] рассматривается послойное построение модели с отдельным критерием оптимизации для каждого слоя. В работах [?, ?, ?] предлагается декомпозиция модели на порождающую и разделяющую, оптимизируемые последовательно. В работе [?] предлагается метод автоматического построения сети, основанный на бустинге. В качестве оптимизируемого функционала предлагается линейная комбинация функции правдоподобия выборки и сложности модели по Радемахеру. В работах [?, ?, ?, ?] предлагается метод автоматического построения сверточной сети с использованием обучения с подкреплением. В [?] используется схожее представление сверточной сети, вместо обучения с подкреплением используется градиентная оптимизация параметров, задающих структуру нейронной сети.

В качестве критерия выбора модели в ряде работ [?, ?, ?, ?, ?, ?] выступает обоснованность модели. В работах [?, ?, ?, ?] рассматривается проблема выбора модели и оценки гиперпараметров в задачах регрессии. Альтернативным критерием выбора модели является минимальная длина описания [?], являю-

щаяся показателем статистической сложности модели и заданной выборки. В работе [?] рассматривается перечень критериев сложности моделей глубокого обучения и их взаимосвязь. В работе [?] в качестве критерия сложности модели выступает показатель нелинейности, характеризуемый степенью полинома Чебышева, аппроксимирующего функцию. В работе [?] анализируется показатель избыточности параметров сети. Утверждается, что по небольшому набору параметров в глубокой сети с большим количеством избыточных параметров можно спрогнозировать значения остальных. В работе [?] рассматривается показатель робастности моделей, а также его взаимосвязь с топологией выборки и классами функций, в частности рассматривается влияние функции ошибки и ее липшицевой константы на робастность моделей. Схожие идеи были рассмотрены в работе [?], в которой исследуется устойчивость классификации модели под действием шума.

Одним из методов получения приближенного значения обоснованности является вариационный метод получения нижней оценки интеграла [?]. В работе [?] рассматривается стохастическая версия вариационного метода. В работе [?] рассматривается алгоритм получения вариационной нижней оценки обоснованности для оптимизации гиперпараметров моделей глубокого обучения. В работе [?] рассматривается получение вариационной нижней оценки интеграла с использованием модификации методов Монте-Карло. В работе [?] рассматривается стохастический градиентный спуск в качестве оператора, порождающего распределение, аппроксимирующее апостериорное распределение параметров модели. Схожий подход рассматривается в работе [?], где также рассматривается стохастический градиентный спуск в качестве оператора, порождающего апостериорное распределение параметров. В работе [?] предлагается модификация стохастического градиентного спуска, аппроксимирующая апостериорное распределение.

Задачей, связанной с проблемой выбора модели, является задача оптимизации гиперпараметров [?, ?]. В работе [?] рассматривается оптимизация гиперпараметров с использованием метода скользящего контроля и методов оптимизации обоснованности моделей, отмечается низкая скорость сходимости гиперпараметров при использовании метода скользящего контроля. В ряде работ [?, ?] рассматриваются градиентные методы оптимизации гиперпараметров, позволяющие оптимизировать большое количество гиперпараметров одновременно. В работе [?] предлагается метод оптимизации гиперпараметров с использованием градиентного спуска с моментом, в качестве оптимизируемого функционала рассматривается ошибка на валидационной части выборки. В работе [?] предлагается метод аппроксимации градиента функции потерь по гиперпараметрам, позволяющий использовать градиентные методы в задаче оптимизации гиперпараметров на больших выборках. В работе [?] предлагается упрощенный метод оптимизации гиперпараметров с градиентным спуском: вместо всей истории обновлений параметров для оптимизации используется только последнее обновление. В работе [?] рассматривается задача оптимизации параметров гра-

диентного спуска с использованием нижней вариационной оценки обоснованности.

Цели работы.

1. Исследовать методы построения моделей глубокого обучения оптимальной и субоптимальной сложности.
2. Предложить критерии оптимальной и субоптимальной сложности модели глубокого обучения.
3. Предложить метод выбора субоптимальной структуры модели глубокого обучения.
4. Предложить алгоритм построения модели субоптимальной сложности и оптимизации ее параметров.

Методы исследования. Для достижения поставленных целей используются методы вариационного байесовского вывода [?, ?, ?]. Рассматривается графовое представление нейронной сети [?, ?]. Для получения вариационных оценок обоснованности модели используется метод, основанный на градиентном спуске [?, ?]. В качестве метода получения модели субоптимальной сложности используется метод автоматического определения релевантности параметров [?, ?] с использованием градиентных методов оптимизации гиперпараметров [?, ?, ?, ?].

Основные положения, выносимые на защиту.

1. Предложен метод байесовского выбора оптимальной и субоптимальной структуры модели глубокого обучения с использованием автоматического определения релевантности параметров.
2. Предложены критерии оптимальной и субоптимальной сложности модели глубокого обучения.
3. Предложен метод графового описания моделей глубокого обучения.
4. Предложено обобщение задачи оптимизации структуры модели, включающее ранее описанные методы выбора модели: оптимизация обоснованности модели, последовательное увеличение сложности модели, последовательное снижение сложности модели, полный перебор вариантов структуры модели.
5. Предложен метод оптимизации вариационной оценки обоснованности модели на основе метода мультистарта задачи оптимизации.
6. Предложен алгоритм оптимизации параметров, гиперпараметров и структурных параметров моделей глубокого обучения.
7. Исследованы свойства оптимизационной задачи при различных значениях метапараметров. Рассмотрены ее асимптотические свойства.

Научная новизна. Разработан новый подход к построению моделей глубокого обучения. Предложены критерии субоптимальной и оптимальной сложности модели, а также исследована их связь. Предложен метод построения модели глубокого обучения субоптимальной сложности. Исследованы методы оптимизации гиперпараметров и параметров модели. Предложена обобщенная задача выбора модели глубокого обучения.

Теоретическая значимость. В целом, данная диссертационная работа носит теоретический характер. В работе предлагаются критерии субоптимальной и оптимальной сложности, основанные на принципе минимальной длины описания. Исследуется взаимосвязь критериев оптимальной и субоптимальной сложности. Предлагаются градиентные методы для получения оценок сложности модели. Доказывается теорема об оценке энтропии эмпирического распределения параметров модели, полученных под действием оператора оптимизации. Доказывается теорема об обобщенной задаче выбора модели глубокого обучения.

Практическая значимость. Предложенные в работе методы предназначены для построения моделей глубокого обучения в прикладных задачах регрессии и классификации; оптимизации гиперпараметров полученной модели; выбора модели из конечного множества заданных моделей; получения оценок переобучения модели.

Степень достоверности и апробация работы. Достоверность результатов подтверждена математическими доказательствами, экспериментальной проверкой полученных методов на реальных задачах выбора моделей глубокого обучения; публикациями результатов исследования в рецензируемых научных изданиях, в том числе рекомендованных ВАК. Результаты работы докладывались и обсуждались на следующих научных конференциях.

1. “Восстановление панельной матрицы и ранжирующей модели в разнородных шкалах”, Всероссийская конференция «57-я научная конференция МФТИ», 2014.
2. “A monolingual approach to detection of text reuse in Russian-English collection”, Международная конференция «Artificial Intelligence and Natural Language Conference», 2015 [**?**].
3. “Выбор модели глубокого обучения субоптимальной сложности с использованием вариационной оценки правдоподобия”, Международная конференция «Интеллектуализация обработки информации», 2016 [**?**].
4. “Machine-Translated Text Detection in a Collection of Russian Scientific Papers”, Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог-21», 2017 [**?**].
5. “Author Masking using Sequence-to-Sequence Models”, Международная конференция «Conference and Labs of the Evaluation Forum», 2017 [**?**].

6. “Градиентные методы оптимизации гиперпараметров моделей глубокого обучения”, Всероссийская конференция «Математические методы распознавания образов ММРО», 2017 [**?**].
7. “Детектирование переводных заимствований в текстах научных статей из журналов, входящих в РИНЦ”, Всероссийская конференция «Математические методы распознавания образов ММРО», 2017 [**?**].
8. “ParaPlagDet: The system of paraphrased plagiarism detection”, Международная конференция «Big Scholar at conference on knowledge discovery and data mining», 2018.
9. “Байесовский выбор наиболее правдоподобной структуры модели глубокого обучения”, Международная конференция «Интеллектуализация обработки информации», 2018 [**?**].
10. “Variational learning across domains with triplet information”, Международная конференция «Visually Grounded Interaction and Language workshop, Conference on Neural Information Processing Systems», 2018.

Публикации по теме диссертации. Основные результаты по теме диссертации изложены в 11 печатных изданиях, 9 из которых изданы в журналах, рекомендованных ВАК.

Личный вклад. Все приведенные результаты, кроме отдельно оговоренных случаев, получены диссертантом лично при научном руководстве д.ф.-м.н. В. В. Стрижова.

Структура и объем работы. Диссертация состоит из оглавления, введения, четырех разделов, заключения, списка иллюстраций, списка таблиц, перечня основных обозначений и списка литературы из 162 наименований. Основной текст занимает 144 страницы.

Основное содержание работы

Во **введении** обоснована актуальность диссертационной работы, сформулированы цели и методы исследования, поставлены основные задачи, обоснована научная новизна, теоретическая и практическая значимость полученных результатов. В **главе 1** приводится формальная постановка задачи выбора модели глубокого обучения. Вводятся основные определения и обозначения, функции качества модели глубокого обучения, описывается вероятностная интерпретация модели.

Проблема выбора структуры модели глубокого обучения формулируется следующим образом: решается задача классификации или регрессии на заданной или пополняемой выборке $\mathfrak{D}$. Требуется выбрать структуру нейронной сети, доставляющей минимум ошибки на этой функции и максимум качества на некотором внешнем критерии. Под моделью глубокого обучения понимается

суперпозиция дифференцируемых по параметрам нелинейный функций. Под структурой модели понимается значения структурных параметров модели, т.е. величин, задающих вид итоговой суперпозиции.

Формализуем описанную выше задачу.

Определение 1. *Объектом* назовем пару $(\mathbf{x}, y), \mathbf{x} \in \mathbb{X} = \mathbb{R}^n, y \in \mathbb{Y}$. В случае задачи классификации $\mathbb{Y}$ является распределением вероятностей принадлежности объекта $\mathbf{x} \in \mathbb{X}$ множеству классов $\{1, \ldots, R\}$: $\mathbb{Y} \subset [0,1]^R$, где R — число классов. В случае задачи регрессии $\mathbb{Y}$ является некоторым подмножеством вещественных чисел $y \in \mathbb{Y} \subseteq \mathbb{R}$. Объект состоит из двух частей: $\mathbf{x}$ соответствует *признаковому описанию объекта*, y — *метке объекта*.

Задана простая выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i = 1, \ldots, m, \tag{1}$$

состоящая из множества объектов

$$\mathbf{x}_i \in \mathbf{X} \subset \mathbb{X}, \quad y_i \in \mathbf{y} \subset \mathbb{Y}.$$

Определение 2. *Моделью* $\mathbf{f}(\mathbf{w}, \mathbf{x})$ назовем дифференцируемую по параметрам $\mathbf{w}$ функцию из множества признаковых описаний объекта во множество меток:

$$\mathbf{f} : \mathbb{W} \times \mathbb{X} \to \mathbb{Y},$$

где $\mathbb{W}$ — пространство параметров функции $\mathbf{f}$.

Специфика задачи выбора модели *глубокого обучения* заключается в том, что модели глубокого обучения могут иметь значительное число параметров, что приводит к неприменимости ряда методов оптимизации и выбора модели. Перейдем к формальному описанию параметрического семейства моделей глубокого обучения.

Определение 3. Пусть задан ациклический граф (V, E), такой что

1. для каждого ребра $(j,k) \in E$: вектор базовых дифференцируемых функций $\mathbf{g}^{j,k} = [\mathbf{g}_0^{j,k}, \ldots, \mathbf{g}_{K^{j,k}-1}^{j,k}]$ мощности $K^{j,k}$;
2. для каждой вершины $v \in V$: дифференцируемая функция агрегации $\mathbf{agg}_v$.
3. Функция $\mathbf{f} = \mathbf{f}_{|V|-1}$, задаваемая по правилу

$$\mathbf{f}_k(\mathbf{w}, \mathbf{x}) = \mathbf{agg}_k \left(\{ \langle \boldsymbol{\gamma}^{j,k}, \mathbf{g}^{j,k} \rangle \circ \mathbf{f}_j(\mathbf{x}) | j \in \mathrm{Adj}(v_k) \} \right), \tag{2}$$

$$k \in \{1, \ldots, |V|-1\}, \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x}, \quad v_k \in V.$$

и являющаяся функцией из признакового пространства $\mathbb{X}$ в пространство меток $\mathbb{Y}$ при значениях векторов, $\boldsymbol{\gamma}^{j,k} \in [0,1]^{K^{j,k}}$.

Граф (V, E) со множеством векторов базовых функций $\{\mathbf{g}^{j,k}, (j,k) \in E\}$ и функций агрегаций $\{\mathbf{agg}_k\}$, где $k \in \{0, \ldots, |V|-1\}$, назовем *параметрическим семейством моделей* $\mathfrak{F}$.

Примером функций агрегации выступают функции суммы и конкатенации векторов.

Определение 4. Функции $\mathbf{f}_0, \ldots, \mathbf{f}_{|V|-1}$ из (2) назовем *слоями или подмоделями* модели $\mathbf{f}$.

Утверждение 1. Для любого значения $\boldsymbol{\gamma}^{j,k} \in [0,1]^{K^{j,k}}$ функция $\mathbf{f} \in \mathfrak{F}$ является моделью.

Определение 5. *Параметрами* модели $\mathbf{f}$ из параметрического семейства моделей $\mathfrak{F}$ назовем конкатенацию векторов параметров всех базовых функций $\{\mathbf{g}^{j,k} | (j,k) \in E\}, \mathbf{w} \in \mathbb{W}$. Вектор параметров базовой функции $\mathbf{g}_l^{j,k}$ будем обозначать как $\mathbf{w}_l^{j,k}$.

Определение 6. Структурой $\boldsymbol{\Gamma}$ модели $\mathbf{f}$ из параметрического семейства моделей $\mathfrak{F}$ назовем конкатенацию векторов $\boldsymbol{\gamma}^{j,k}$. Множество всех возможных значений структуры $\boldsymbol{\Gamma}$ будем обозначать как $\mathbb{\Gamma}$. Векторы $\boldsymbol{\gamma}^{j,k}, (j,k) \in E$ назовем *структурными параметрами модели.*

Определение 7. *Параметризацией* множества моделей M назовем параметрическое семейство моделей $\mathfrak{F}$, такое что для каждой модели $\mathbf{f} \in M$ существуют значение структуры модели $\boldsymbol{\Gamma}$ при котором функция $\mathbf{f}$ совпадает с функцией (2).

Рассмотрим варианты ограничений, которые накладываются на структурные параметры $\boldsymbol{\gamma}^{j,k}$ параметрического семейства моделей. Цель данных ограничений — уточнение архитектуры модели глубокого обучения, которую требуется получить.

1. Структурные параметры лежат на веришнах булевого куба: $\boldsymbol{\gamma}^{j,k} \in \{0,1\}^{K^{j,k}}$. Структурные параметры $\boldsymbol{\gamma}^{j,k}$ интерпретируются как параметр включения или выключения компонент вектора базовых функций $\mathbf{g}^{j,k}$ в итоговую модель.
2. Структурные параметры лежат внутри булевого куба: $\boldsymbol{\gamma} \in [0,1]^{K^{j,k}}$. Релаксированная версия предыдущих ограничений, позволяющая проводить градиентную оптимизацию для структурных параметров.
3. Структурные параметры лежат на веришнах симплекса: $\boldsymbol{\gamma}^{j,k} \in \bar{\Delta}^{K^{j,k}-1}$. Каждый вектор структурных параметров $\boldsymbol{\gamma}^{j,k}$ имеет только одну ненулевую компоненту, определяющую какая из базовых функций $\mathbf{g}^{j,k}$ войдет в итоговую модель. Примером параметрического семейства моделей, требующим такое ограничение является семейство полносвязанных нейронных сетей с одним скрытым слоем и двумя значениями количества нейронов на скрытом слое. Схема семейства представлена на Рис. **??**. Данное семейство можно представить как семейство с двумя базовыми функциями вида $\mathbf{g} = \boldsymbol{\sigma}(\mathbf{w}^\mathsf{T}\mathbf{x})$, где матрицы параметров каждой из функций $\mathbf{g}^{1,1}, \mathbf{g}^{1,2}$ имеют фиксированное число нулевых столбцов. Количество этих столбцов определяет размерность итогового скрытого пространства или числа нейронов на скрытом слое.

4. Структурные параметры лежат внутри симплекса: $\boldsymbol{\gamma}^{j,k} \in \Delta^{K^{j,k}-1}$. Релаксированная версия предыдущих ограничений, позволяющая проводить градиентную оптимизацию для структурных параметров. Значение стуктурных параметров $\boldsymbol{\gamma}^{j,k}$ интерпретируются как вклад каждой компоненты вектора базовых функций $\mathbf{g}^{j,k}$ в итоговую модель.

В данной работе рассматривается случай, когда на структурные параметры наложено ограничение 4. Данные ограничения позволяют решать задачу выбора модели как для семейства моделей типа многослойных полносвязных нейронных сетей, так и для более сложных параметрических семейств [**?**].

Для дальнейшей постановки задачи введем понятие вероятностной модели, и связанных с ним определений. Будем полагать, что для параметров модели $\mathbf{w}$ и структуры $\boldsymbol{\Gamma}$ задано распределение $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$, соответствующее предположениям о распределении структуры и параметров.

Определение 8. *Гиперпараметрами* $\mathbf{h} \in \mathbb{H}$ модели назовем параметры распределения $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$.

Определение 9. *Априорным распределением* параметров и структуры модели назовем вероятностное распределение, соответствующее предположениям о распределении параметров модели:

$$p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) : \mathbb{W} \times \mathbb{\Gamma} \to \mathbb{R}^+,$$

где $\mathbb{W}$ — множество значений параметров модели, $\mathbb{\Gamma}$ — множество значений структуры модели.Формальное определение метапараметров $\boldsymbol{\lambda} \in \mathbb{\Lambda}$ будет дано далее.

Одной из постановок задачи выбора структуры модели является *двусвязный байесовский вывод.* На *первом уровне* байесовского вывода находится апостериорное распределение параметров.

Определение 10. *Апостериорным распределением* назовем распределение вида

$$p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})}{p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})} \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}). \quad (3)$$

Определение 11. *Вероятностной моделью глубокого обучения* назовем совместное распределение вида

$$p(\mathbf{y}, \mathbf{w}, \boldsymbol{\Gamma}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}, \boldsymbol{\lambda}) p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) : \mathbb{Y}^m \times \mathbb{W} \times \mathbb{\Gamma} \to \mathbb{R}^+.$$

Определение 12. *Функцией правдоподбия выборки* назовем величину

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) : \mathbb{Y}^m \to \mathbb{R}^+.$$

На *втором уровне* байесовского вывода осуществляется выбор модели на основе обоснованности модели.

Определение 13. *Обоснованностью модели* назовем величину

$$p(\mathbf{y}|\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = \iint_{\mathbf{w},\boldsymbol{\Gamma}} p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma})p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})d\mathbf{w}d\boldsymbol{\Gamma}. \quad (4)$$

Получение значений апостериорного распределения и обоснованности модели сетей глубокого обучения является вычислительно сложной процедурой. Для получения оценок на данные величины используют методы, такие как аппроксимация Лапласа [?] и вариационная нижняя оценка [?]. В данной работе в качестве метода получения оценок обоснованности модели выступает вариационная нижняя оценка.

Определение 14. *Вариационным распределением* назовем параметрическое распределение $q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})$, являющееся приближением апостериорного распределения параметров и структуры $p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$.

Определение 15. *Вариационными параметрами* модели $\boldsymbol{\theta} \in \Theta$ назовем параметры вариационного распределения $q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})$.

Определение 16. Пусть задано вариационное распределения $q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})$. *Функцией потерь* $L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$ для модели $\mathbf{f}$ назовем дифференцируемую функцию, принимаемую за качество модели на обучающей выборки при параметрах модели, получаемых из распределения q.

В качестве функции $L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$ может выступать логарифм правдоподобия выборки $\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma})$ и логарифм апостериорной вероятности $\log p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$ параметров и структуры модели на обучающей выборке.

Определение 17. Пусть задано вариационное распределения $q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})$ и функция потерь $L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$. *Функцией валидации* $Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda})$ для модели $\mathbf{f}$ назовем дифференцируемую функцию, принимаемую за качество модели при векторе $\boldsymbol{\theta}$, заданном неявно.

В данной работе задача выбора структуры модели и параметров модели ставится как двухуровневая задача оптимизации:

$$\mathbf{h}^* = \arg\max_{\mathbf{h}\in\mathbb{H}} Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}^*,\boldsymbol{\lambda}), \quad (5)$$

где $\boldsymbol{\theta}^*$ — решение задачи оптимизации

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}). \quad (6)$$

Определение 18. *Задачей выбора модели* $\mathbf{f}$ назовем двухуровневую задачу оптимизации (5),(6).

Методы, используемые для оптимизации гиперпараметров моделей глубокого обучения должны быть эффективными по вычислительным затратам в силу высокой вычислительной сложности оптимизации параметров модели. В [?, ?] рассматривается задача оптимизации гиперпараметров стохастическими методами. В [?] проводится сравнение случайного поиска значений гиперпараметров с переборным алгоритмом. В [?] производится сравнение случайного поиска и алгоритмов, основанных на вероятностных моделях.

Градиентные методы оптимизации гиперпараметров.

Определение 19. Назовем *оператором оптимизации* алгоритм T выбора вектора параметров $\boldsymbol{\theta}'$ по параметрам предыдущего шага $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}' = T(\boldsymbol{\theta}|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}), \tag{7}$$

где $\boldsymbol{\lambda}$ — параметры оператора оптимизации или *метапараметры*.

Метапараметры соответствуют параметрам оптимизации, т.е. параметрам, которые не подлежат оптимизации в ходе задачи выбора модели.

Пример схожего описания оптимизации модели с использованием оператора оптимизации можно найти в [**?**].

Частным случаем оператора оптимизации является оператор стохастического спуска:

$$T(\boldsymbol{\theta}|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \boldsymbol{\theta} - \lambda_{\text{lr}} \nabla(-L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})), \tag{8}$$

где λ_{lr} — шаг градиентного спуска, $\hat{\mathbf{y}}, \hat{\mathbf{X}}$ — случайная подвыборка заданной мощности выборки $\mathfrak{D}$.

В случае оптимизации гиперпараметров оператор оптимизации применяется не к вариационным параметрам $\boldsymbol{\theta}$, а к гиперпараметрам $\mathbf{h}$:

$$\mathbf{h} = T(\mathbf{h}|Q, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}). \tag{9}$$

В случае, если для решения задачи (6) применяется несколько шагов оператора оптимизации (7), $\boldsymbol{\theta}^*$ рассматривается как рекурсивная функция от начального приближения вариационных параметров $\boldsymbol{\theta}^0$ и вектора гиперпараметров $\mathbf{h}$:

$$\boldsymbol{\theta}^* = T \circ \cdots \circ T(\boldsymbol{\theta}|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \boldsymbol{\theta}^*(\boldsymbol{\theta}^0, \mathbf{h}). \tag{10}$$

Решение задачи оптимизации (9) при (10) является вычислительно сложным, поэтому применяются методы, аппроксимирующие применение градиентных методов при (10). В **главе 2** рассматривается задача выбора моделей глубокого обучения субоптимальной сложности. Вводятся вероятностные предположения о распределении параметров. В качестве сложности модели выступает *обоснованность модели* (4). Для получения оценки обоснованности применяются вариационные методы с использованием градиентных алгоритмов оптимизации. Предполагается, что структура $\boldsymbol{\Gamma}$ модели глубокого обучения $\mathbf{f}$ и метапараметры $\boldsymbol{\lambda}$ определены однозначно:

$$p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) = p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}), \quad p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) = p(\mathbf{w}|\mathbf{h}), \quad p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) = p(\mathbf{y}|\mathbf{X}, \mathbf{w}).$$

Определение 20. Сложностью модели $\mathbf{f}$ назовем обоснованность модели:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{h}) = \int_{\mathbf{w} \in \mathbb{W}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\mathbf{h}) d\mathbf{w}. \tag{11}$$

Определение 21. Модель $\mathbf{f}$ назовем оптимальной среди моделей M, если достигается максимум интеграла (11).

Требуется найти оптимальную модель $\mathbf{f}$ из заданного множества моделей M, а также значения ее параметров $\mathbf{w}$, доставляющие максимум апостериорной вероятности

$$p(\mathbf{w}|\mathbf{y},\mathbf{X},\mathbf{h}) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathbf{y}|\mathbf{X},\mathbf{h})}. \tag{12}$$

Интеграл обоснованности (11) модели является трудновычислимым для данного семейства моделей. Одним из методов вычисления приближенного значения обоснованности является получение вариационной оценки обоснованности.

Определение 22. Вариационной оценкой логарифма обоснованности модели (11) $\log p(\mathbf{y}|\mathbf{X},\mathbf{h})$ называется оценка $\log \hat{p}(\mathbf{y}|\mathbf{X},\mathbf{h})$, полученная аппроксимацией неизвестного апостериорного распределения $p(\mathbf{w}|\mathbf{y},\mathbf{X},\mathbf{h})$ заданным распределением $q(\mathbf{w})$.

Будем рассматривать задачу нахождения вариационной оценки как задачу оптимизации. Пусть задано множество распределений $\mathfrak{Q} = \{q(\mathbf{w})\}$. Сведем задачу нахождения наиболее близкой вариационной нижней оценки интеграла (11) к оптимизации вида

$$\hat{q}(\mathbf{w}) = \arg\max_{q\in\mathfrak{Q}} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{h})}{q(\mathbf{w})} d\mathbf{w}.$$

В данной работе в качестве множества $\mathfrak{Q}$ рассматривается нормальное распределение и распределение параметров, неявно получаемое оптимизацией градиентными методами.

Оценка (**??**) является нижней, поэтому может давать некорректные оценки для обоснованности (11). Для того, чтобы оценить величину этой ошибки, докажем следующее утверждение.

Теорема 1 ([**?**])**.** Пусть задано множество $\mathfrak{Q} = \{q(\mathbf{w})\}$ непрерывных распределений. Максимизация вариационной нижней оценки

$$\int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{h})}{q(\mathbf{w})} d\mathbf{w}$$

логарифма интеграла (11) эквивалентна минимизации расстояния Кульбака–Лейблера между распределением $q(\mathbf{w}) \in \mathfrak{Q}$ и апостериорным распределением параметров $p(\mathbf{w}|\mathbf{y},\mathbf{X},\mathbf{h})$:

$$\hat{q} = \arg\max_{q\in\mathfrak{Q}} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{h})}{q(\mathbf{w})} d\mathbf{w} \Leftrightarrow \hat{q} = \arg\min_{q\in\mathfrak{Q}} \mathrm{D}_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y},\mathbf{X},\mathbf{h})\big), \tag{13}$$

$$\mathrm{D}_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y},\mathbf{X},\mathbf{h})\big) = \int_{\mathbf{w}} q(\mathbf{w}) \log \left(\frac{q(\mathbf{w})}{p(\mathbf{w}|\mathbf{y},\mathbf{X},\mathbf{h})}\right) d\mathbf{w}.$$

Определение 23. Пусть задано множество распределений $\mathfrak{Q}$. Модель $\mathbf{f}$ назовем субоптимальной на множестве моделей M, если модель доставляет максимум нижней вариационной оценке интеграла (13)

$$\max_{q\in\mathfrak{Q}} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{h})}{q(\mathbf{w})} d\mathbf{w}. \tag{14}$$

В качестве множества $\mathfrak{Q} = \{q(\mathbf{w})\}$ рассматривается параметрическое семейство нормальных распределений с диагональными матрицами ковариаций:

$$q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}), \quad \boldsymbol{\theta} = [\boldsymbol{\mu}_q, \mathbf{diag}(\mathbf{A}_q^{-1})] \tag{15}$$

где $\mathbf{A}_q$ — диагональная матрица ковариаций, $\boldsymbol{\mu}_q$ — вектор средних компонент.

В качестве множества распределений $\mathfrak{Q} = \{q(\mathbf{w})\}$, аппроксимирующих неизвестное распределение $\log p(\mathbf{y}|\mathbf{X},\mathbf{h})$, используются распределения параметров, полученные в ходе их оптимизации.

Представим неравенство (**??**)

$$\log p(\mathbf{y}|\mathbf{X},\mathbf{h}) \geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{h})}{q(\mathbf{w})} d\mathbf{w} = \mathsf{E}_{q(\mathbf{w})} \log p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{h}) - \mathsf{S}\big(q(\mathbf{w})\big), \tag{16}$$

где S — энтропия распределения:

$$\mathsf{S}\big(q(\mathbf{w})\big) = -\int_{\mathbf{w}} q(\mathbf{w}) \log q(\mathbf{w}) d\mathbf{w},$$

$$p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{h}) = p(\mathbf{w}|\mathbf{h}) p(\mathbf{y}|\mathbf{X},\mathbf{w}),$$

$\mathsf{E}_{q(\mathbf{w})} \log p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{h})$ — матожидание логарифма вероятности $\log p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{h})$:

$$\mathsf{E}_{q(\mathbf{w})} \log p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{h}) = \int_{\mathbf{w}} \log p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{h}) q(\mathbf{w}) d\mathbf{w}.$$

Оценка распределений производится при оптимизации параметров. Оптимизация выполняется в режиме мультистарта [**?**], т.е. при запуске оптимизации параметров модели из нескольких разных начальных приближений. Основная проблема такого подхода — вычисление энтропии S распределений $q(\mathbf{w}) \in \mathfrak{Q}$. Ниже представлен метод получения оценок энтропии (20) S и оценок обоснованности (16).

Запустим r процедур оптимизаций модели $\mathbf{f}$ из разных начальных приближений:

$$L(\boldsymbol{\theta}|\mathbf{h},\mathbf{X},\mathbf{y}) = \sum_{l=1}^{r} \log p(\mathbf{y},\mathbf{w}^l|\mathbf{X},\mathbf{h}) \to \max, \quad \boldsymbol{\theta} = [\mathbf{w}^1,\ldots,\mathbf{w}^r],$$

где r — число оптимизаций,

$$\log p(\mathbf{y}, \mathbf{w}^l|\mathbf{X}, \mathbf{h}) = \sum_{i=1}^{m} \log p(y_i, \mathbf{w}^l|\mathbf{x}_i, \mathbf{h}) = \log\, p(\mathbf{w}^l|\mathbf{h}) + \sum_{i=1}^{m} \log p(y_i|\mathbf{x}_i, \mathbf{w}^l, \mathbf{h}). \tag{17}$$

Пусть начальные приближения параметров $\mathbf{w}^1, \ldots, \mathbf{w}^r$ порождены из некоторого начального распределения $q^0(\mathbf{w})$:

$$\mathbf{w}^1, \ldots, \mathbf{w}^r \sim q^0(\mathbf{w}).$$

Для дальнейшего описания метода введем понятие оператора градиентного спуска, являющегося частным случаем оператора оптимизации (7).

Определение 24. Оператором градиентного спуска назовем оператор оптимизации вида

$$T(\boldsymbol{\theta}|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \boldsymbol{\theta} - \lambda_{\text{lr}} \nabla(-L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})), \tag{18}$$

где λ_{lr} — длина шага градиентного спуска.

В данной главе будем рассматривать распределения, полученные из нескольких точек старта оптимизации параметров $\mathbf{w}$ модели $\mathbf{f}$. Для удобства будем использовать $L(\mathbf{w})$ как эквивалентную форму записи $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ для $\boldsymbol{\theta} = [\mathbf{w}]^{\text{T}}$, и $T(\mathbf{w})$ как эквивалентную форму записи $T(\boldsymbol{\theta}|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$.

Теорема 2. Пусть T — оператор градиентного спуска, L — функция потерь, градиент ∇L которой имеет константу Липшица C_L. Пусть $\boldsymbol{\theta} = [\mathbf{w}^1, \ldots, \mathbf{w}^r]^{\text{T}}$ — начальные приближения оптимизации модели, где r — число начальных приближений. Пусть λ_{lr} — длина шага градиентного спуска, такая что

$$\lambda_{\text{lr}} < \frac{1}{C_L}, \quad \lambda_{\text{lr}} < \big(\max_{l \in \{1,\ldots,r\}} \lambda_{\max}(\mathbf{H}(\mathbf{w}^l))\big)^{-1}, \tag{19}$$

где $\lambda_{\max}$ — наибольшее по модулю собственное значение гессиана $\mathbf{H}$ минус функции потерь $(-L)$.

При выполнении неравенств (19) разность энтропий распределений $q'(\mathbf{w}), q(\mathbf{w})$ на смежных шагах почти наверное сходится к следующему выражению:

$$\mathsf{S}\big(q'(\mathbf{w})\big) - \mathsf{S}\big(q(\mathbf{w})\big) \approx \frac{1}{r} \sum_{l=1}^{r} \big(-\lambda_{\text{lr}} \text{Tr}[\mathbf{H}(\mathbf{w}'^l)] - \lambda_{\text{lr}} \text{Tr}[\mathbf{H}(\mathbf{w}'^l)\mathbf{H}(\mathbf{w}'^l)]\big) + o_{\lambda_{\text{lr}}^2 \to 0}(1). \tag{20}$$

Теорема 3. Оценка (16) на шаге оптимизации τ представима в виде

$$\log \hat{p}(\mathbf{y}|\mathbf{X}, \mathbf{h}) \approx \frac{1}{r} \sum_{g=1}^{r} L(\mathbf{w}_\tau^l|\mathbf{X}, \mathbf{y}) + \tag{21}$$

$$+\mathsf{S}\big(q^0(\mathbf{w})\big) + \frac{1}{r} \sum_{b=1}^{\tau} \sum_{l=1}^{r} \big(-\lambda_{\text{lr}} \text{Tr}[\mathbf{H}(\mathbf{w}_b^l)] - \lambda_{\text{lr}}^2 \text{Tr}[\mathbf{H}(\mathbf{w}_b^l)\mathbf{H}(\mathbf{w}_b^l)]\big)$$

с точностью до слагаемых вида $o_{\lambda_{\text{lr}}^2 \to 0}(1)$, где $\mathbf{w}_b^l$ — l-я реализация параметров модели на шаге оптимизации b, $q^0(\mathbf{w})$ — начальное распределение.

В **главе 3** рассматривается задача оптимизации гиперпараметров модели глубокого обучения. Для оптимизации гиперпараметров модели предлагаются алгоритмы, основанные на градиентном спуске. Так как сложность рассматриваемых алгоритмах сопоставима со сложностью оптимизации параметров модели, предлагается оптимизировать параметры и гиперпараметры в единой процедуре. Для выбора адекватных значений гиперпараметров вводятся вероятностные предположения о распределении параметров. В качестве оптимизируемой функции выступает байесовская обоснованность модели и кросс-валидация. Как и в предыдущей главе, предполагается, что структура модели $\mathbf{\Gamma}$ для вероятностной модели глубокого обучения $\mathbf{f}$ и метапараметры $\boldsymbol{\lambda}$ определены однозначно:

$$p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) = p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}), \quad p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) = p(\mathbf{w}|\mathbf{h}), \quad p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) = p(\mathbf{y}|\mathbf{X}, \mathbf{w}).$$

Пусть априорное распределение параметров имеет вид

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}), \tag{22}$$

где $\mathbf{A}^{-1} = \text{diag}[\alpha_1, \ldots, \alpha_u]^{-1}$ — матрица ковариаций диагонального вида, где u — количество параметров $\mathbf{w}$ модели $\mathbf{f}$.

требуется найти параметры $\boldsymbol{\theta}^*$ и гиперпараметры $\mathbf{h}^*$ модели, доставляющие максимум следующей функции:

$$\mathbf{h}^* = \arg\max_{\mathbf{h} \in \mathbb{H}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}), \tag{23}$$

$$\boldsymbol{\theta}^*(\mathbf{h}) = \arg\max_{\boldsymbol{\theta} \in \mathbb{\Theta}} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}), \tag{24}$$

где L, Q — функции потерь и валидации.

Рассмотрим вид переменной $\boldsymbol{\theta}$ и функций L, Q для различных методов выбора модели и оптимизации ее параметров.

Базовый метод. Пусть оптимизация параметров и гиперпараметров производится по всей выборке $\mathfrak{D}$ по одной и той же функции $L = Q$:

$$L(\boldsymbol{\theta}|\mathbf{h}, \mathbf{X}, \mathbf{y}) = Q(\mathbf{h}|\boldsymbol{\theta}, \mathbf{X}, \mathbf{y}) = \log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{h}) = \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) + \log p(\mathbf{w}|\mathbf{h}).$$

Вариационным параметрам $\boldsymbol{\theta}$ модели $\mathbf{f}$ соответствует вектор параметров модели:

$$\boldsymbol{\theta} = \mathbf{w}.$$

Кросс-валидация. Разобьем выборку $\mathfrak{D}$ случайно на K равных частей:

$$\mathfrak{D} = \mathfrak{D}_1 \sqcup \cdots \sqcup \mathfrak{D}_k, \mathfrak{D}_k = \{\mathbf{X}_k, \mathbf{y}_k\}, \quad k = 1, \ldots, K.$$

Запустим K оптимизаций модели, каждую на своей части выборки. Положим $\boldsymbol{\theta} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$, где $\mathbf{w}_1, \ldots, \mathbf{w}_K$ — параметры модели при оптимизациях $1, \ldots, K$.

Положим функцию L пропорциональной среднему значению логарифма апостериорной вероятности по всем $k-1$ разбиениям $\mathfrak{D}$:

$$L = \frac{1}{K}\sum_{k=1}^{K}\Big(\frac{K}{K-1}\log p(\mathbf{y}_k|\mathbf{X}_k, \mathbf{w}_k) + \log p(\mathbf{w}_k|\mathbf{h})\Big). \tag{25}$$

Положим функцию Q равной среднему значению правдоподобия выборки по частям выборки $\mathfrak{D}_k$, на которых не проходила оптимизация параметров:

$$Q = \frac{1}{k}\sum_{q=1}^{k} k\log p(\mathbf{y} \setminus \mathbf{y}_k|\mathbf{X}_k \setminus \mathbf{X}, \mathbf{w}_q).$$

где операция «$\mathbf{X} \setminus \mathbf{X}_k$» определяется как взятие описаний всех объектов $\mathbf{X}$ за исключением описаний объектов из $\mathbf{X}_k$.

Вариационная оценка обоснованности. Положим $L = Q$, равной вариационной оценке обоснованности модели:

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{h}) \geq -\mathrm{D}_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\big) + \int_{\mathbf{w}} q(\mathbf{w})\log p(\mathbf{y}|\mathbf{X}, \mathbf{w})d\mathbf{w} \approx \tag{26}$$

$$\approx \sum_{i=1}^{m}\log p(y_i|\mathbf{x}_i, \mathbf{w}_i) - D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\big) = L = Q,$$

где q — нормальное распределение с диагональной матрицей ковариаций:

$$q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}), \tag{27}$$

где $\mathbf{A}_q = \mathrm{diag}[\alpha_1^q, \ldots, \alpha_u^q]^{-1}$ — диагональная матрица ковариаций, $\boldsymbol{\mu}_q$ — вектор средних компонент. Расстояние D_{KL} между двумя гауссовыми величинами задается как

$$D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\big) = \frac{1}{2}\big(\mathrm{Tr}[\mathbf{A}\mathbf{A}_q^{-1}] + (\boldsymbol{\mu}-\boldsymbol{\mu}_q)^{\mathsf{T}}\mathbf{A}(\boldsymbol{\mu}-\boldsymbol{\mu}_q) - u + \ln|\mathbf{A}^{-1}| - \ln|\mathbf{A}_q^{-1}|\big).$$

В качестве вариационных параметров $\boldsymbol{\theta}$ выступают параметры распределения q:

$$\boldsymbol{\theta} = [\alpha_1, \ldots, \alpha_u, \mu_1, \ldots, \mu_u].$$

0.1. Градиентные методы оптимизации гиперпараметров

В данном разделе приводится описание рассматриваемых градиентных методов. Краткая характеристика и основные преимущества каждого из представленных методов отображены в Табл. 1, 2.

Алгоритм	**Тип алгоритма**	**Преимущества алгоритма**	**Недостатки алгоритма**
Случайный поиск	стохастический	простота реализации	Алгоритм неэффективен при большом количестве гиперпараметров (проклятие размерности)
Жадный алгоритм [?]	градиентный	Возможность одновременной оптимизации параметров и гиперпараметров	Жадность алгоритма
HOAG [?]	градиентный	Быстрая сходимость	Алгоритм требователен к настройкам параметров
DrMAD [?]	градиентный	Алгоритм учитывает алгоритм оптимизации параметров модели и его параметры	Алгоритм страдает от проблем неустойчивости градиентного спуска (градиентный взрыв и затухание); Алгоритм работает в строгих предположениях о линейности траектории оптимизации гиперпараметров.

Таблица 1. Преимущества и недостатки рассматриваемых алгоритмов

Алгоритм	**Тип**	**Сложность итерации оптимизации**	**Предположения**
Случайный поиск	стохастический	$O(\eta\|\Theta\| \cdot \|\hat{\mathfrak{D}}\|)$	-
Жадный алгоритм [?]	градиентный	$O(\eta\|\Theta\| \cdot \|\mathbb{H}\| \cdot \|\hat{\mathfrak{D}}\|)$	$\mathbf{H}(\boldsymbol{\theta}) = \mathbf{I}$
HOAG [?]	градиентный	$O(\eta\|\Theta\| \cdot \|\mathbb{H}\| \cdot \|\hat{\mathfrak{D}}\| + o)$, где o —сложность решения системы линейных уравнений	Первая производная Q и вторая производная L являются липшицевыми функциями $\det\mathbf{H} \neq \mathbf{0}$;
DrMAD [?]	градиентный	$O(\eta\|\Theta\| \cdot \|\mathbb{H}\| \cdot \|\hat{\mathfrak{D}}\|)$	Траектория оптимизации вариационных параметров $\boldsymbol{\theta} = \boldsymbol{\theta}^0, \dots \boldsymbol{\theta}^\eta$ линейна

Таблица 2. Сложность и предположения для различных алгоритмов оптимизации гиперпараметров

В **главе 4** рассматривается задача выбора структуры модели глубокого обучения. Предлагается ввести вероятностные предположения о распределении

параметров и распределении структуры модели. В качестве оптимизируемой функции для гиперпараметров модели предлагается обобщенная функция ее обоснованности. Показывается, что данная функция оптимизирует ряд критериев выбора структуры модели: метод максимального правдоподобия, последовательное увеличение и снижению сложности модели, полный перебор структуры модели, а также получение максимума вариационной оценки обоснованности модели.

Определим априорные распределения параметров и структуры модели следующим образом. Пусть для каждого ребра $(j,k) \in E$ и каждой базовой функции $\mathbf{g}_l^{j,k}$ параметры модели $\mathbf{w}_l^{j,k}$ распределены нормально с нулевым средним:

$$\mathbf{w}_l^{j,k} \sim \mathcal{N}\big(\mathbf{0}, (\gamma_l^{j,k})^2 (\mathbf{A}_l^{j,k})^{-1}\big),$$

где $(\mathbf{A}_l^{j,k})^{-1}$ — диагональная матрица, $l \in \{0, \ldots, K^{j,k}-1\}$, где $K^{j,k}$ — количество базовых функций для ребра $K^{j,k}$. Априорное распределение $p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h})$ параметров $\mathbf{w}_l^{j,k}$ зависит не только от гиперпараметров $\mathbf{A}_k^{j,k}$, но и от структурного параметра $\gamma_l^{j,k} \in (0,1)$.

В качестве априорного распределения для структуры $\boldsymbol{\Gamma}$ предлагается использовать произведение распределений Gumbel-Softmax ($\mathcal{GS}$) [**?**]:

$$p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) = \prod_{(j,k)\in E} p(\boldsymbol{\gamma}^{j,k}|\mathbf{s}^{j,k}, \lambda_{\text{temp}}),$$

где для каждого структурного параметра $\boldsymbol{\gamma}^{j,k}$ с количеством базовых функций $K^{j,k}$ вероятность $p(\boldsymbol{\gamma}^{j,k}|\mathbf{s}^{j,k}, \lambda_{\text{temp}})$ определена следующим образом:

$$p(\boldsymbol{\gamma}^{j,k}|\mathbf{s}^{j,k}, \lambda_{\text{temp}}) = (K^{j,k}-1)!(\lambda_{\text{temp}})^{K^{h,j}-1} \prod_{l=0}^{K^{j,k}-1} s_l^{j,k} (\boldsymbol{\gamma}_l^{j,k})^{-\lambda_{\text{temp}}-1} \times \\ \times \left(\sum_{l=0}^{K^{j,k}-1} s_l^{j,k} (\boldsymbol{\gamma}_l^{j,k})^{-\lambda_{\text{temp}}} \right)^{-K^{j,k}}, \quad (28)$$

где $\mathbf{s}^{j,k} \in (0,\infty)^{Kj,k}$ — гиперпараметр, отвечающий за смещенность плотности распределения относительно точек симплекса на $K^{j,k}$ вершинах, $\lambda_{\text{temp}} > 0$ — метапараметр температуры, отвечающий за концентрацию плотности вблизи вершин симплекса или в центре симплекса.

TODO: GS, репараметризация

В качестве регуляризатора для матрицы $(\mathbf{A}_l^{j,k})^{-1}$ предлагается использовать обратное гамма-распределение:

$$(\mathbf{A}_l^{j,k})^{-1} \sim \text{inv-gamma}(\lambda_1, \lambda_2),$$

где $\lambda_1, \lambda_2 \in \boldsymbol{\lambda}$ — метапараметры оптимизации.

Таким образом, предлагаемая вероятностная модель содержит следующие компоненты:

1. Параметры $\mathbf{w}$ модели, распределенные нормально.
2. Структура модели $\mathbf{\Gamma}$, содержащая все структурные параметры $\{\boldsymbol{\gamma}^{j,k}, (j,k) \in E\}$, распределенные по распределению Gumbel-Softmax.
3. Гиперпараметры $\mathbf{h} = [\mathrm{diag}(\mathbf{A}), \mathbf{s}]$, где $\mathbf{A}$ — конкатенация матриц $\mathbf{A}^{j,k}, (j,k) \in E$, $\mathbf{s}$ — конкатенация параметров Gumbel-Softmax распределений $\mathbf{s}^{j,k}, (j,k) \in E$, где E — множество ребер, соответствующих графу рассматриваемого параметрического семейства моделей $\mathfrak{F}$.
4. Метапараметры: $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_{\text{temp}}]$. Эти параметры не подлежат оптимизации и задаются экспертно.

В качестве критерия выбора гиперпараметров предлагается использовать апостериорную вероятность гиперпараметров:

$$p(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\lambda}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) p(\mathbf{h}|\boldsymbol{\lambda}) \to \max_{\mathbf{h} \in \mathbb{H}}. \tag{29}$$

Структура модели и параметры модели выбираются на основе полученных значений гиперпараметров:

$$\mathbf{w}^*, \mathbf{\Gamma}^* = \arg\max_{\mathbf{w} \in \mathbb{W}, \mathbf{\Gamma} \in \mathbb{\Gamma}} p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}^*, \boldsymbol{\lambda}),$$

где $\mathbf{h}^*$ — решение задачи оптимизации (**??**).

Для вычисления обоснованности модели

$$p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \iint_{\mathbf{\Gamma}, \mathbf{w}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) p(\mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) d\mathbf{\Gamma} d\mathbf{w}$$

из (**??**) предлагается использовать нижнюю вариационную оценку обоснованности.

Теорема 4. Пусть $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}) = q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ — вариационное распределение с параметрами $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\mathbf{w}}, \boldsymbol{\theta}_{\mathbf{\Gamma}}]$, аппроксимирующее апостериорное распределение структуры и параметров:

$$q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}) \approx p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}),$$

$$q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) \approx p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}),$$

$$q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) \approx p(\mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}).$$

Тогда справедлива следующая оценка:

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) \geq \tag{30}$$

$$\mathsf{E}_{q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - D_{\mathrm{KL}}\big(q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) || p(\mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) -$$

$$- D_{\mathrm{KL}}\big(q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) || p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})\big),$$

где $D_{\mathrm{KL}}\big(q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})||p(\mathbf{w}|\boldsymbol{\Gamma},\mathbf{h},\boldsymbol{\lambda})\big)$ вычисляется по формуле условной дивергенции:

$$D_{\mathrm{KL}}\big(q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})||p(\mathbf{w}|\boldsymbol{\Gamma},\mathbf{h},\boldsymbol{\lambda})\big) = \mathsf{E}_{\boldsymbol{\Gamma}\sim q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})}\mathsf{E}_{\mathbf{w}\sim q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}\log\left(\frac{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}{p(\mathbf{w}|\boldsymbol{\Gamma},\mathbf{h},\boldsymbol{\lambda})}\right).$$

Для анализа сложности полученной модели введем понятие *параметрической сложности.*

Определение 25. Параметрической сложностью $C_p(\boldsymbol{\theta}|U_{\mathbf{h}},\boldsymbol{\lambda})$ модели с вариационными параметрами $\boldsymbol{\theta}$ на компакте $U_{\mathbf{h}} \subset \mathbb{H}$ назовем минимальную дивергенцию между вариационным и априорным распределением:

$$C_p(\boldsymbol{\theta}|U_{\mathbf{h}},\boldsymbol{\lambda}) = \min_{\mathbf{h}\in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big).$$

Одним из критериев удаления неинформативных параметров в вероятностных моделях является отношение вариационной плотности параметров в нуле к вариационной плотности параметра в моде распределения (**??**):

Определение 26. Относительной вариационной плотностью параметра $w \in \mathbf{w}$ при условии структуры $\boldsymbol{\Gamma}$ и гиперпараметров $\mathbf{h}$ назовем отношение вариационной плотности в моде априорного распределения параметра к вариационной плотности в моде вариационного распределения параметра:

$$\rho(w|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}},\mathbf{h},\boldsymbol{\lambda}) = \frac{q_{\mathbf{w}}(\operatorname{mode} p(w|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})|\boldsymbol{\Gamma},\mathbf{h},\boldsymbol{\lambda})}{q_{\mathbf{w}}(\operatorname{mode} q_{\mathbf{w}}(w|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}.$$

Относительной вариационной плотностью вектора параметров $\mathbf{w}$ назовем следующее выражение:

$$\rho(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}},\mathbf{h},\boldsymbol{\lambda}) = \prod_{w\in\mathbf{w}} \rho(w|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}},\mathbf{h},\boldsymbol{\lambda}). \tag{31}$$

Теорема 5. Пусть выполнены условия Леммы **??** и Леммы **??**. Тогда справедливо следующее выражение:

$$\lim_{i\to\infty} \mathsf{E}_{q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i])}\rho(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}[i],\mathbf{h}[i],\boldsymbol{\lambda})^{-1} = 1.$$

Рассмотрим основные статистические критерии выбора вероятностных моделей.

1. Критерий максимального правдоподобия:

$$\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) \to \max_{\mathbf{w}\in U_{\mathbf{w}},\boldsymbol{\Gamma}\in U_{\boldsymbol{\Gamma}}}.$$

   Для использования данного критерия в качестве задачи выбора модели предлагается следующее обобщение:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})}\log\ p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}). \tag{32}$$

Данное обобщение (**??**) эквивалентно критерию максимального правдоподобия при выборе в качестве $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$ распределения, основанного на запуске нескольких оптимизаций параметров (17) и структуры. Метод не предполагает оптимизации гиперпараметров $\mathbf{h}$. Для формального соответствия данной задачи задаче выбора модели (5),(6), т.е. двухуровневой задачи оптимизации, положим $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) \to \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}},$$

$$Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) \to \max_{\mathbf{h} \in U_{\mathbf{h}}}.$$

2. Метод максимальной апостериорной вероятности.

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) \to \max_{\mathbf{w} \in U_{\mathbf{w}}, \boldsymbol{\Gamma} \in U_{\boldsymbol{\Gamma}}}.$$

Аналогично предыдущему методу сформулируем вариационное обобщение данной задачи:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \tag{33}$$

$$= \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \big(\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) + \log p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) \big).$$

Т.к. в рамках данной задачи (**??**) не предполагается оптимизации гиперпараметров $\mathbf{h}$, положим параметры распределения $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ фиксированными:

$$\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_{\text{temp}}, \mathbf{s}, \text{diag}(\mathbf{A})].$$

3. Полный перебор структуры:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) = p'|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) \tag{34}$$

где p' — некоторое распределение на структуре $\boldsymbol{\Gamma}$, выступающее в качестве метапараметра.

4. Критерий Акаике:

$$\text{AIC} = 2 \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - 2|\mathbb{W}| \to \max.$$

В случае, если рассматриваемые модели принадлежат одному параметрическому семейству моделей $\mathfrak{F}$, то количество параметров у всех рассматриваемых моделей совпадает. Тогда критерий Акаике совпадает с критерием максимального правдоподобия. Для использования критерия Акаике для сравнения моделей, принадлежащих одному параметрическому семейству $\mathfrak{F}$ предлагается следующая переформулировка:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \tag{35}$$

$$-|\{w : D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) < \lambda_{\text{prune}}\}|,$$

где

$$\mathbf{h} = \arg\min_{\mathbf{h}' \in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big), \tag{36}$$

λ_{prune} — метапараметр алгоритма, $U_{\mathbf{h}} \subset \mathbb{H}$ — область определения задачи по гиперпараметрам. Предложенное обобщение (**??**) применимо только в случае, если выражение (**??**) определено однозначно, т.е. существует единственный вектор гиперпараметров $\mathbf{h} \in U_{\mathbf{h}}$, доставляющий минимум дивергенции $D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big)$.

5. Информационный критерий Шварца:

$$\mathrm{BIC} = 2\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - |\mathbb{W}|\log m \to \max.$$

Переформулируем данный критерий аналогично критерию AIC:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) = \tag{37}$$

$\log \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - 0.5\log m|\{w : D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) < \lambda_{\text{prune}}\}|$,

метапараметр λ_{prune} определен аналогично (**??**).

6. Метод вариационной оценки обоснованности:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = \tag{38}$$

$$= \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) + \\ + \log p(\mathbf{h}|\boldsymbol{\lambda}) \to \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}},$$

$$Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) =$$

$$= \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) + \\ + \log p(\mathbf{h}|\boldsymbol{\lambda}) \to \max_{\mathbf{h} \in U_{\mathbf{h}}},$$

В рамках данной задачи функции $L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$ и $Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda})$ совпадают, все гиперпараметры $\mathbf{h}$ подлежат оптимизации.

7. Валидация на отложенной выборке:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}},\mathbf{w},\boldsymbol{\Gamma}) + \log p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda}) \to \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}}, \tag{39}$$

$$Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}},\mathbf{w},\boldsymbol{\Gamma}) \to \max_{\mathbf{h} \in U_{\mathbf{h}}},$$

где $(\mathbf{X}_{\text{train}},\mathbf{y}_{\text{train}}), (\mathbf{X}_{\text{test}},\mathbf{y}_{\text{test}})$ — разбиение выборки на обучающую и контрольную подвыборку. В рамках данной задачи, все гиперпараметры $\mathbf{h}$ подлежат оптимизации.

Определение 27. Двухуровневую задачу оптимизации будем называть *обобщающей* на компакте

$$U = U_{\boldsymbol{\theta}_{\mathbf{w}}} \times U_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}} \times U_{\mathbf{h}} \times U_{\boldsymbol{\lambda}} \subset \Theta_{\mathbf{w}} \times \Theta_{\boldsymbol{\Gamma}} \times \mathbb{H} \times \mathbb{\Lambda},$$

если она удовлетворяет следующим критериям.

1. Область определения каждого параметра $w \in \mathbf{w}$, гиперпараметра $h \in \mathbf{h}$ и метапараметра $\lambda \in \boldsymbol{\lambda}$ не является пустым множеством и не является точкой.
2. Для каждого значения гиперпараметров $\mathbf{h}$ оптимальное решение нижней задачи оптимизации (6)

$$\boldsymbol{\theta}^*(\mathbf{h}) = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$$

определено однозначно при любых значениях метапараметров $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$.
3. Критерий максимизации правдоподобия выборки: существует $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$ и $K_1 > 0$,

$$K_1 < \max_{\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}} Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2), \boldsymbol{\lambda}),$$

такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}$, удовлетворяющих неравенству

$$Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2), \boldsymbol{\lambda}) > K_1,$$

выполняется неравенство

$$\mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}^*(\mathbf{h}_1))} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) > \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}^*(\mathbf{h}_2))} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}).$$

4. Критерий минимизации параметрической сложности: существует $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$ и $K_2 > 0$,

$$K_2 < \max_{\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}} Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2), \boldsymbol{\lambda}),$$

такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}$, удовлетворяющих неравенству

$$Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}_2), \boldsymbol{\lambda}) > K_2,$$

параметрическая сложность первой модели меньше, чем второй:

$$C_p(\boldsymbol{\theta}^*(\mathbf{h}_1)|U_{\mathbf{h}}, \boldsymbol{\lambda}) < C_p(\boldsymbol{\theta}^*(h_2)|U_{\mathbf{h}}, \boldsymbol{\lambda}).$$

5. Критерий приближения оценки обоснованности: существует значение гиперпараметров $\boldsymbol{\lambda}$, такое что значение функций потерь $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ как сложной функции от $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ пропорционально вариационной оценки обоснованности модели:

$$Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*(\mathbf{h}), \boldsymbol{\lambda}) \propto$$

$$\propto \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}'(\mathbf{h}))} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}'(\mathbf{h}))||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) + \log p(\mathbf{h}|\boldsymbol{\lambda})$$

для всех $\mathbf{h} \in U_{\mathbf{h}}$, где в качестве гиперпараметров $\mathbf{h}$ рассматриваются все гиперпараметры модели, вне зависимости от критерия и особенности его оптимизации гиперпараметров:

$$\mathbf{h} = [\mathbf{A}, \mathbf{s}],$$

где

$$\boldsymbol{\theta}'(\mathbf{h}) = \arg\max_{\boldsymbol{\theta} \in U_{\mathbf{h}}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big).$$

6. Критерий перебора оптимальных структур: существует константа $K_3 > 0$, такая что существует хотя бы одна пара гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}$, удовлетворяющая неравенствам:

$$D_{\mathrm{KL}}\big(p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})||p(\boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\big) > K_3, D_{\mathrm{KL}}\big(p(\boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})||p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})\big) > K_3$$

и набор метапараметров $\boldsymbol{\lambda}$, такие что для произвольных локальных оптимумов $\mathbf{h}_1, \mathbf{h}_2$ задачи оптимизации $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$, полученных при метапараметрах $\boldsymbol{\lambda}$ и удовлетворяющих неравенствам

$$D_{\mathrm{KL}}\big(p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})||p(\boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\big) > K_3, D_{\mathrm{KL}}\big(p(\boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})||p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})\big) > K_3,$$

$$Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) > Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}),$$

существует значение метапараметров $\boldsymbol{\lambda}' \neq \boldsymbol{\lambda}$, такие что
   (a) соответствие между вариационными параметрами $\boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\theta}^*(\mathbf{h}_2)$ сохраняется при $\boldsymbol{\lambda}'$,
   (b) выполняется неравенство $Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}') < Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}')$.
7. Критерий непрерывности: функции $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ и $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ непрерывны по метапараметрам $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$.

Первый критерий является техническим и используется для исключения из рассмотрения вырожденных задач оптимизации. Второй критерий говорит о том, что решение первого и второго уровня должны быть согласованы и определены однозначно. Критерии 3-5 определяют возможные критерии оптимизации, которые должны приближаться обобщающей задачей. Критерий 6 говорит о возможности перехода между различными структурами модели. Данный

критерий говорит о том, что мы можем перейти от одного набора гиперпараметров $\mathbf{h}_1$ к другим $\mathbf{h}_2$, если они соответствуют локальным оптимумам задачи оптимизации, и дивергенция соответствующих априорных распределений на структурах $p(\mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ значимо высока. При этом соответствующие вариационные распределения $q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ могут оказаться достаточно близки, несмотря на значимые различия априорных распределений. Поэтому возможным дополнением этого критерия был бы критерий, позволяющий переходить от структуры к структуре, если соответствующие распределения $q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ различаются значимо. Последний критерий говорит о том, что обобщающая задача должна позволять производить переход между различными методами выбора параметров и структуры модели непрерывно.

Теорема 6. Рассмотренные задачи (**??**),(**??**),(**??**),(**??**),(**??**),(**??**) не являются обобщающими.

Теорема 7. Пусть $q_{\mathbf{\Gamma}}$ — абсолютно непрерывное распределение с дифференцируемой плотностью, такой что:

1. Градиент плотности $\nabla_{\boldsymbol{\theta}_{\mathbf{\Gamma}}} q(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ является ненулевым почти всюду.
2. Выражение $\nabla_{\boldsymbol{\theta}_{\mathbf{\Gamma}}} q(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) \log\, p(\mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ ограничено на $U_{\boldsymbol{\theta}}$ абсолютно непрерывной случайной величиной, не зависящей от $\mathbf{\Gamma}$, с конечным первым моментом.

Тогда задача (**??**) не является обобщающей.

В качестве обобщающей задачи оптимизации предлагается оптимизационную задачу следующего вида:

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \tag{40}$$

$$\begin{aligned} &= \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} \mathsf{E}_{q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta}^*)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - \\ &- \lambda^{\mathrm{Q}}_{\mathrm{prior}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}^*)||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) - \\ &- \sum_{p' \in \mathfrak{P}, \lambda \in \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}} \lambda D_{\mathrm{KL}}\big(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}^*)||p'\big) + \log p(\mathbf{h}|\boldsymbol{\lambda}), \end{aligned}$$

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \tag{41}$$

$$= \mathsf{E}_{q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - \lambda^{\mathrm{L}}_{\mathrm{prior}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}^*)||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big),$$

где $\mathfrak{P}$ — непустое множество распределений на структуре $\mathbf{\Gamma}$, $\lambda^{\mathrm{Q}}_{\mathrm{prior}}, \lambda^{\mathrm{L}}_{\mathrm{prior}}, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$ — некоторые числа. Множество распределений $\mathfrak{P}$ отвечает за перебор структур $\mathbf{\Gamma}$ в процессе оптимизации модели. В предельном случае, когда температура λ_{temp} близка к нулю, а множество $\mathfrak{P}$ состоит из распределений, близких к дискретным, соответствующим всем возможным структурам, калибровка $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$ порождает последовательность задач оптимизаций, схожую с перебором структур. Рассмотрим следующий пример.

Теорема 8. Пусть

1. Задан компакт $U = U_{\boldsymbol{\theta}_\mathbf{w}} \times U_{\boldsymbol{\theta}_\boldsymbol{\Gamma}} \times U_\mathbf{h} \times U_{\boldsymbol{\lambda}}$, где априорное распределение $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ и распределение $p(\mathbf{h}|\boldsymbol{\lambda})$ непрерывны на $U_\mathbf{h} \times U_{\boldsymbol{\lambda}}$.
2. Задано непустое множество $\mathfrak{P}$ абсолютно непрерывных распределений на структуре, чьи плотности непрерывны и не принимают нулевое значение, где хотя бы одно распределение $p_1 \in \mathfrak{P}$ является Gumbel-Softmax распределением, и для каждого значения $\mathbf{s} \in U_\mathbf{h}, \lambda_{\text{temp}} \in U_{\boldsymbol{\lambda}}$, существует значение параметров распределения p_1, такое что $p_1 = p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$. Параметры распределений $p \in \mathfrak{P}$ принадлежат множеству метапараметров $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$.
3. Вариационное распределение $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$ является абсолютно непрерывным, плотность которого непрерывна по метапараметрам $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$ и не принимает нулевое значение.
4. Область определения каждого параметра $w \in \mathbf{w}$, гиперпараметра $h \in \mathbf{h}$ и метапараметра $\lambda \in \boldsymbol{\lambda}$ не является пустым и не является точкой.
5. Для каждого значения гиперпараметров $\mathbf{h} \in U_\mathbf{h}$ оптимальное решение нижней задачи оптимизации $\boldsymbol{\theta}^*$ определено однозначно на $U_{\boldsymbol{\theta}} = U_{\boldsymbol{\theta}_\mathbf{w}} \times U_{\boldsymbol{\theta}_\boldsymbol{\Gamma}}$ при любых значениях метапараметров $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$.
6. Область значений метапараметров $\lambda^{\text{Q}}_{\text{likelihood}}, \lambda^{\text{Q}}_{\text{prior}}, \lambda^{\text{L}}_{\text{prior}}, \boldsymbol{\lambda}^{\text{Q}}_{\text{struct}}$ включает отрезок от нуля до единицы.
7. Существует значение метапараметров

$$\lambda_1 > 0, \lambda_2 > 0, \lambda^{\text{Q}}_{\text{likelihood}} > 0 \in U_{\boldsymbol{\lambda}},$$

такое что

$$\max_{\mathbf{h} \in U_\mathbf{h}} \log p(\mathbf{h}|\boldsymbol{\lambda}) - \min_{\mathbf{h} \in U_\mathbf{h}} \log p(\mathbf{h}|\boldsymbol{\lambda}) < \max_{\mathbf{h} \in U_\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\mathbf{h} \in U_\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$$

при $\boldsymbol{\lambda}^{\text{Q}}_{\text{struct}} = \mathbf{0}, \lambda^{\text{Q}}_{\text{prior}} = 0$.
8. Существует значение метапараметров

$$\lambda^{\text{L}}_{\text{prior}} > 0, \lambda^{\text{Q}}_{\text{prior}} > 0, \lambda_1 > 0, \lambda_2 > 0, \lambda_{\text{temp}} > 0 \in U_{\boldsymbol{\lambda}},$$

такое что

$$\max_{\mathbf{h} \in U_\mathbf{h}} \frac{1}{\lambda^{\text{Q}}_{\text{prior}}} \log p(\mathbf{h}|\boldsymbol{\lambda}) - \min_{\mathbf{h} \in U_\mathbf{h}} \frac{1}{\lambda^{\text{Q}}_{\text{prior}}} \log p(\mathbf{h}|\boldsymbol{\lambda}) +$$

$$+ \max_{\mathbf{h} \in U_\mathbf{h}} \min_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}} D_{\text{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) -$$

$$- \min_{\mathbf{h} \in U_\mathbf{h}, \boldsymbol{\theta} \in U_{\boldsymbol{\theta}}} D_{\text{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) + \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}} \frac{1}{\lambda^{\text{L}}_{\text{prior}}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) -$$

$$- \min_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}} \frac{1}{\lambda^{\text{L}}_{\text{prior}}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) <$$

$$< \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}, \mathbf{h} \in U_\mathbf{h}} D_{\text{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) -$$

$$- \min_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}, \mathbf{h} \in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big)$$

при $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}, \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 0$.

9. Существуют значения метапараметров $\lambda^{\mathrm{Q}}_{\mathrm{prior}} > 0, \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} > 0, \lambda_1 > 0, \lambda_2 > 0, \lambda_{\mathrm{temp}} > 0 \in U_{\boldsymbol{\lambda}}$, такие что существуют гиперпараметры $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}$:

$$D_{\mathrm{KL}}\big(p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda}) || p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\big) > \\ > \frac{\max_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})}{m_\lambda},$$

$$D_{\mathrm{KL}}\big(p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}) || p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})\big) > \\ > \frac{\max_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})}{m_\lambda}$$

при $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$, где m_λ — максимальное значение $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$ перед распределением p_1 из первого условия теоремы.

Тогда задача (**??**) является обобщающей на U.

Следующие теоремы говорят о соответствии предлагаемой обобщающей задачи вероятностной модели. В частности, задача оптимизации параметров и гиперпараметров соответствует двухуровневому байесовскому выводу.

Теорема 9. Пусть $\lambda^{\mathrm{Q}}_{\mathrm{prior}} = \lambda^{\mathrm{L}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 1, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$. Тогда:

1. Задача оптимизации (**??**) доставляет максимум апостериорной вероятности гиперпараметров с использованием вариационной оценки обоснованности:

$$\mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) + \\ + \log p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) \to \max_{\mathbf{h}}.$$

2. Вариационное распределение $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$ приближает апостериорное распределение $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ наилучшим образом:

$$D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})\big) \to \min_{\boldsymbol{\theta}}.$$

3. Если существуют такие значения параметров $\boldsymbol{\theta}_{\mathbf{w}}, \boldsymbol{\theta}_{\boldsymbol{\Gamma}}$, что $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) = q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}), p(\boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$, то решение задачи оптимизации $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ доставляет эти значения вариационных параметров.

Докажем, что варьирование коэффициента $\lambda^{\mathrm{L}}_{\mathrm{prior}}$ приводит к оптимизации вариационной оценки обоснованности для выборки из той же генеральной совокупности, но другой мощности.

Теорема 10. Пусть $m \gg 0$, $\lambda^{\mathrm{L}}_{\mathrm{prior}} > 0$, $\frac{m}{\lambda^{\mathrm{L}}_{\mathrm{prior}}} \in \mathbb{N}$, $\frac{m}{\lambda^{\mathrm{L}}_{\mathrm{prior}}} \gg 0$. Тогда оптимизация функции

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \lambda^{\mathrm{L}}_{\mathrm{prior}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big)$$

эквивалентна оптимизации вариационной оценки обоснованности

$$\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\hat{\mathbf{y}}|\hat{\mathbf{X}}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big)$$

для произвольной случайной подвыборки $\hat{\mathbf{y}}, \hat{\mathbf{X}}$ мощности $\frac{m}{\lambda^{\mathrm{L}}_{\mathrm{prior}}}$ из генеральной совокупности.

Теорема 11. Пусть

1. Выполнены условия Леммы **??**.
2. Функция $Q(\mathbf{h}|\boldsymbol{\theta}_2, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ является вогнутой по $\mathbf{h} \in U_{\mathbf{h}}$ при $\lambda^{\mathrm{Q}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$.
3. Решение задачи (**??**) единственно при $\lambda^{\mathrm{Q}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$.
4. Все стационарные точки $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$ функции $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ являются решениями нижней задачи оптимизации при $\lambda^{\mathrm{Q}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$ с обратимым гессианом.
5. Значения $p(\mathbf{h}|\boldsymbol{\lambda})$ приблизительно равны на $U_{\mathbf{h}}$:

$$p(\mathbf{h}_1|\boldsymbol{\lambda}) \approx p(\mathbf{h}_2|\boldsymbol{\lambda}) \text{ для всех } \mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}.$$

Тогда справедлива следующая оценка разности параметрических сложностей:

$$C_p(\boldsymbol{\theta}_1|U_{\mathbf{h}}, \boldsymbol{\lambda}_1) - C_p(\boldsymbol{\theta}_2|U_{\mathbf{h}}, \boldsymbol{\lambda}_2) < \frac{\lambda^{\mathrm{L}}_{\mathrm{prior}}}{\lambda^{\mathrm{Q}}_{\mathrm{prior}}}(\lambda^{\mathrm{Q}}_{\mathrm{prior}_2} - \lambda^{\mathrm{L}}_{\mathrm{prior}}) \times$$

$$\times \max_{\mathbf{h}\in U_{\mathbf{h}}, \boldsymbol{\theta}\in U_{\boldsymbol{\theta}}} \nabla_{\boldsymbol{\theta},\mathbf{h}}(D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||[p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})]\big))^{\mathsf{T}} \nabla^2_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}_2))^{-1} \times$$

$$\times \nabla_{\boldsymbol{\theta}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) \max_{\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}} ||\mathbf{h}_1 - \mathbf{h}_2||.$$

Теорема 12. Пусть $\frac{\lambda^{\mathrm{Q}}_{\mathrm{prior}}}{\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}} = \lambda^{\mathrm{L}}_{\mathrm{prior}}$. Тогда задача оптимизации (**??**) представима в виде одноуровневой задачи оптимизации:

$$\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \lambda^{\mathrm{Q}}_{\mathrm{prior}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) -$$

$$- \sum_{p'\in\mathfrak{P}, \lambda\in\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}} D_{\mathrm{KL}}\big(p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})||p'\big) - \log p(\mathbf{h}|\boldsymbol{\lambda}) \to \max_{\mathbf{h},\boldsymbol{\theta}}.$$

В **главе 5** продемонстрировано применение предложенных методов к прикладным задачам классификации и регрессии, определения схожести предложений на основе их векторных представлений, а также к задачам прореживания моделей глубокого обучения.

В заключении представлены основные результаты диссертационной работы.

1. Предложен метод байесовского выбора оптимальной и субоптимальной структуры модели глубокого обучения с использованием автоматического определения релевантности параметров.
2. Предложены критерии оптимальной и субоптимальной сложности модели глубокого обучения.
3. Предложен метод графового описания моделей глубокого обучения. Предложено обобщение задачи оптимизации структуры модели, включающее ранее описанные методы выбора модели: оптимизация обоснованности модели, последовательное увеличение сложности модели, последовательное снижение сложности модели, полный перебор вариантов структуры модели.
4. Предложен метод оптимизации вариационной оценки обоснованности модели на основе метода мультистарта задачи оптимизации.
5. Предложен алгоритм оптимизации параметров, гиперпараметров и структурных параметров моделей глубокого обучения.
6. Исследованы свойства оптимизационной задачи при различных значениях метапараметров. Рассмотрены ее асимптотические свойства.

Публикации соискателя по теме диссертации

Публикации в журналах из списка ВАК.

1. Бахтеев О.Ю., Попова М.С., Стрижов В.В., "Системы и средства глубокого обучения в задачах классификации", Системы и средства информатики, 26:2 (2016), 4–22 [**?**].
2. Bakhteev, O., Kuznetsova, R., Romanov, A. and Khritankov, A., 2015, November. A monolingual approach to detection of text reuse in Russian-English collection. In 2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT) (pp. 3-10). IEEE [**?**].
3. Romanov, A., Kuznetsova, R., Bakhteev, O. and Khritankov, A., 2016. Machine-Translated Text Detection in a Collection of Russian Scientific Papers. Computational Linguistics and Intellectual Technologies. 2016 [**?**].
4. Bakhteev, O. and Khazov, A., 2017. Author Masking using Sequence-to-Sequence Models. In CLEF (Working Notes). 2017 [**?**].
5. Бахтеев О.Ю., Стрижов В.В., "Выбор моделей глубокого обучения субоптимальной сложности", Автоматика и телемеханика, 2018, № 8, 129–147; Automation Remote Control, 79:8 (2018), 1474–1488 [**?**].
6. Огальцов А.В., Бахтеев О.Ю., "Автоматическое извлечение метаданных из научных PDF-документов", Информатика и её применения, 12:2 (2018), 75–82 [**?**].

7. Смердов А.Н., Бахтеев О.Ю., Стрижов В.В., “Выбор оптимальной модели рекуррентной сети в задачах поиска парафраза”, Информатика и её применения, 12:4 (2018), 63–69 [?].
8. Грабовой А.В., Бахтеев О.Ю., Стрижов В.В. “Определение релевантности параметров нейросети”, Информатика и её применения. 13:2 (2019), 62-71 [?].
9. Bakhteev, O.Y. and Strijov, V.V., 2019. Comprehensive analysis of gradient-based hyperparameter optimization algorithms. Annals of Operations Research, pp.1-15 [?].

Остальные публикации.

10. Бахтеев О.Ю. Восстановление панельной матрицы и ранжирующей модели по метризованной выборке в разнородных данных. // Машинное обучение и анализ данных. 2016. № 7. С. 72-77 [?].
11. Бахтеев О.Ю. Восстановление пропущенных значений в разнородных шкалах с большим числом пропусков. // Машинное обучение и анализ данных. 2015. № 11. С. 1-11 [?].