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: **Стрижов Вадим Викторович**
доктор физико-математических наук, Федеральный исследовательский центр «Информатика и управление» Российской академии наук, отдел интеллектуальных систем, ведущий научный сотрудник.

Официальные оппоненты: **Чуличков Алексей Иванович**
доктор физико-математических наук, профессор, Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М. В. Ломоносова», профессор кафедры математического моделирования и информатики физического факультета.

Зайцев Алексей Алексеевич
кандидат физико-математических наук, Автономная некоммерческая образовательная организация высшего образования «Сколковский институт науки и технологий», руководитель лаборатории в Центре по научным и инженерным вычислительным технологиям для задач с большими массивами данных.

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики».

Защита состоится « » февраля 2020 года в 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 2020 года.

И. о. ученого секретаря
диссертационного совета Д 002.073.05
д.т.н. И. А. Матвеев

Общая характеристика работы

Актуальность темы. В работе исследуется проблема автоматического выбора моделей глубокого обучения оптимальной и субоптимальной сложности. Под сложностью модели понимается *минимальная длина описания* (Grünwald: 2005), минимальное количество информации, которое требуется для передачи информации о модели и о выборке. Получение 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 (Bishop: 2006), основанные на аппроксимации неизвестного апостериорного распределения другим заданным распределением. Под субоптимальной сложностью понимается нижняя вариационная оценка обоснованности модели.

Одна из проблем построения моделей глубокого обучения — большое число параметров моделей (Hinton: 2007, 2013). Поэтому задача выбора моделей глубокого обучения включает в себя выбор вычислительно эффективной стратегии построения модели. В работе (Barron: 2008) приводятся теоретические оценки построения нейросетей с использованием жадных стратегий, при которых построение модели производится итеративно последовательным увеличением числа нейронов в сети. В работах (Zoph: 2016, Baker: 2017, Cai: 2018, Zoph: 2018) предлагаются методы автоматического построения моделей глубокого обучения, основанные на обучении с подкреплением. В (Liu: 2018) предлагается градиентная оптимизация структуры модели.

В качестве критерия выбора модели в ряде работ (MacKay: 2002, Bishop: 2006, Стрижов: 2010, 2014) выступает обоснованность модели. Альтернативными критериями выступают показатель нелинейности модели (Vladislavleva: 2008), робастность модели (Xu: 2012) и эксплуатационные критерии качества модели. Важным свойством, предъявляемым к критериям качества модели, является устойчивость выбранных моделей под действием шума (Szegedy: 2013).

Одним из методов получения приближенного значения обоснованности является вариационный метод получения нижней оценки интеграла (Bishop: 2006). Использование вариационной оценки в качестве приближения обоснованности позволяет аппроксимировать апостериорное распределение с использованием широкого семейства распределений. В работе (Graves: 2011) рассматривается алгоритм получения вариационной нижней оценки обоснованности для оптимизации параметров и гиперпараметров моделей глубокого обучения. В работе (Maclaurin: 2015) рассматривается стохастический градиентный спуск в качестве оператора, порождающего распределение, аппроксимирующее апостериорное распределение параметров модели. Схожий подход предлагается в работе (Mandt: 2017), где также рассматривается стохастический градиентный спуск в качестве оператора, приближающего апостериорное распределение параметров.

Задачей, связанной с проблемой выбора модели, является задача оптимизации гиперпараметров (MacKay: 2002, Bishop: 2006). В работе (Стрижов: 2012) рассматривается оптимизация гиперпараметров с использованием метода скользящего контроля и методов оптимизации обоснованности моделей, отмечается низкая скорость сходимости оптимизации гиперпараметров при использовании метода скользящего контроля. В ряде работ (Maclaurin: 2015, Domke: 2012) рассматриваются градиентные методы оптимизации гиперпараметров, позволяющие оптимизировать большое количество гиперпараметров одновременно. В работе (Maclaurin: 2015) предлагается метод оптимизации гиперпараметров с использованием градиентного спуска с моментом, в качестве оптимизируемой функции рассматривается ошибка на валидационной части выборки. В работах (Pedregosa: 2016, Luketina: 2016) предлагается метод аппроксимации градиента функции потерь по гиперпараметрам, позволяющий использовать градиентные методы в задаче оптимизации гиперпараметров на больших выборках.

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 байесовского вывода. В качестве оценки обоснованности выступает вариационная нижняя оценка обоснованности модели. 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. 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 $\mathfrak{D}, (\mathbf{x}, y) \in \mathfrak{D}, \mathbf{x} \in \mathbb{X} = \mathbb{R}^n, y \in \mathbb{Y}$. 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 значение структурных параметров модели, т.е. величин, задающих вид итоговой суперпозиции.

Определение 1. *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 $\mathbf{f}$.

Определение 2. Пусть задан ациклический граф (V, E), такой, что 1) для каждого ребра $(j, k) \in E$ задан 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 $v \in V$ задана дифференцируемая функция агрегации $\mathbf{agg}_v$; 3) задана функция $\mathbf{f} = \mathbf{f}_{|V|-1}$:

$$\mathbf{f}_k(\mathbf{w}, \mathbf{x}) = \mathbf{agg}_k\left(\{\langle \boldsymbol{\gamma}^{j,k}, \mathbf{g}^{j,k}\rangle \circ \mathbf{f}_j(\mathbf{x}) | j \in \mathrm{Adj}(v_k)\}\right), \tag{1}$$

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

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* $\mathfrak{F}$.

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

Определение 3. 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.*

В работе рассматривается случай, когда структурные параметры лежат внутри симплекса: $\boldsymbol{\gamma}^{j,k} \in \Delta^{K^{j,k}-1}$.

Определение 4. *Гиперпараметрами* $\mathbf{h} \in \mathbb{H}$ модели назовем параметры распределения $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$.

Определение 5. *Априорным распределением* параметров и структуры модели назовем вероятностное распределение, соответствующее предположениям о распределении параметров модели: $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) : \mathbb{W} \times \mathbb{\Gamma} \to \mathbb{R}^+$, где $\mathbb{W}$ — множество значений параметров модели, $\mathbb{\Gamma}$ — множество значений структуры модели.

Определение 6. *Апостериорным распределением* назовем распределение вида

$$p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma})p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})}{p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})} \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma})p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}). \quad (2)$$

Определение 7. *Обоснованностью модели* назовем величину

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

Определение 8. *Вариационным распределением* назовем параметрическое распределение $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})$ с параметрами $\boldsymbol{\theta} \in \mathbb{\Theta}$, являющееся приближением апостериорного распределения параметров и структуры $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$.

Определение 9. Пусть задано вариационное распределения $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})$. *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.

Определение 10. Пусть задано вариационное распределения $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})$ 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.

Задача выбора структуры модели и параметров модели ставится как двухуровневая задача оптимизации:

$$\mathbf{h}^* = \arg\max_{\mathbf{h} \in \mathbb{H}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*, \boldsymbol{\lambda}), \quad \boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \mathbb{\Theta}} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}). \quad (4)$$

Определение 11. *Задачей выбора модели* $\mathbf{f}$ назовем двухуровневую задачу оптимизации (4).

Метапараметры $\boldsymbol{\lambda}$ соответствуют параметрам оптимизации, т.е. параметрам, которые не подлежат оптимизации в ходе задачи выбора модели.

В **главе 2** рассматривается задача выбора моделей глубокого обучения субоптимальной сложности. Вводятся вероятностные предположения о распределении параметров. В качестве сложности модели выступает *обоснованность модели* (3). Для получения оценки обоснованности применяются вариационные методы с использованием градиентных алгоритмов оптимизации. Предполагается, что структура $\mathbf{\Gamma}$ модели глубокого обучения $\mathbf{f}$ и метапараметры $\boldsymbol{\lambda}$ определены однозначно:

$$p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) = p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}), \quad p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) = p(\mathbf{w}|\mathbf{h}), \quad p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) = p(\mathbf{y}|\mathbf{X}, \mathbf{w}). \quad (5)$$

Пусть априорное распределение параметров имеет вид

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}), \tag{6}$$

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

Определение 12. Сложностью модели $\mathbf{f}$ назовем обоснованность модели:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{h}) = \int_{\mathbf{w}\in\mathbb{W}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\mathbf{h}) d\mathbf{w}. \tag{7}$$

Определение 13. Модель $\mathbf{f}$ назовем оптимальной среди моделей множества M, если достигается максимум интеграла (7).

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

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{h}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\mathbf{h})}{p(\mathbf{y}|\mathbf{X}, \mathbf{h})}. \tag{8}$$

В качестве функции, приближающей логарифм интеграла (7), будем рассматривать его вариационную нижнюю оценку, полученную при помощи неравенства Йенсена:

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{h}) \geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{h})}{q(\mathbf{w})} d\mathbf{w} = \tag{9}$$

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

где $\mathrm{D}_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\big)$ — расстояние Кульбака–Лейблера между двумя распределениями.

Определение 14. Пусть задано множество распределений $\mathfrak{Q}$. Модель $\mathbf{f}$ назовем субоптимальной на множестве моделей M, если модель доставляет максимум нижней вариационной оценке интеграла (9).

В качестве множества $\mathfrak{Q}$ рассматривается два семейства распределений. Первое семейство — семейство нормальных распределений с диагональными матрицами ковариаций:

$$q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}), \quad \boldsymbol{\theta} = [\boldsymbol{\mu}_q, \mathbf{diag}(\mathbf{A}_q^{-1})] \tag{10}$$

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

В качестве второго семейства распределений $\mathfrak{Q}$, рассматриваются распределения параметров, полученные в ходе оптимизации модели.

Представим интеграл (9) в виде:

$$\mathsf{E}_{q(\mathbf{w})} \log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{h}) - \mathsf{S}\big(q(\mathbf{w})\big), \tag{11}$$

где S — энтропия распределения:

$$\mathsf{S}\big(q(\mathbf{w})\big) = -\int_{\mathbf{w}} q(\mathbf{w}) \log q(\mathbf{w}) d\mathbf{w}, \quad p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{h}) = p(\mathbf{w}|\mathbf{h}) p(\mathbf{y}|\mathbf{X}, \mathbf{w}).$$

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 (12) S и оценок обоснованности (11).

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

Для удобства будем использовать $L(\mathbf{w})$ как эквивалентную форму записи $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ для $\boldsymbol{\theta} = [\mathbf{w}]^{\mathrm{T}}$.

Определение 15. Оператором градиентного спуска назовем оператор оптимизации вида $T(\mathbf{w}) = \mathbf{w} - \lambda_{\mathrm{lr}} \nabla(-L(\mathbf{w}))$, где λ_{lr} — длина шага градиентного спуска.

Теорема 1. Пусть T — оператор градиентного спуска, L — функция потерь, градиент ∇L которой имеет константу Липшица C_L. Пусть $\boldsymbol{\theta} = [\mathbf{w}^1, \ldots, \mathbf{w}^r]^{\mathrm{T}}$ — начальные приближения оптимизации модели, где r — число начальных приближений. Пусть λ_{lr} — длина шага градиентного спуска, такая, что $\lambda_{\mathrm{lr}} < \frac{1}{C_L}, \quad \lambda_{\mathrm{lr}} < \big(\max_{l \in \{1,\ldots,r\}} \lambda_{\max}(\mathbf{H}(\mathbf{w}^l))\big)^{-1}$, где $\lambda_{\max}$ — наибольшее по модулю собственное значение гессиана $\mathbf{H}$ минус функции потерь $(-L)$.

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

$$\mathsf{S}\big(q'(\mathbf{w})\big) - \mathsf{S}\big(q(\mathbf{w})\big) \approx \frac{1}{r} \sum_{l=1}^{r} \big(-\lambda_{\mathrm{lr}} \mathrm{Tr}[\mathbf{H}(\mathbf{w}'^l)] - \lambda_{\mathrm{lr}} \mathrm{Tr}[\mathbf{H}(\mathbf{w}'^l)\mathbf{H}(\mathbf{w}'^l)]\big) + o_{\lambda_{\mathrm{lr}}^2 \to 0}(1). \tag{12}$$

Теорема 2. Оценка (11) на шаге оптимизации τ представима в виде

$$\frac{1}{r} \sum_{g=1}^{r} L(\mathbf{w}_\tau^l|\mathbf{X}, \mathbf{y}) + \mathsf{S}\big(q^0(\mathbf{w})\big) + \frac{1}{r} \sum_{b=1}^{\tau} \sum_{l=1}^{r} \big(-\lambda_{\mathrm{lr}} \mathrm{Tr}[\mathbf{H}(\mathbf{w}_b^l)] - \lambda_{\mathrm{lr}}^2 \mathrm{Tr}[\mathbf{H}(\mathbf{w}_b^l)\mathbf{H}(\mathbf{w}_b^l)]\big) \tag{13}$$

с точностью до слагаемых вида $o_{\lambda_{\mathrm{lr}}^2 \to 0}(1)$, 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. Предполагается, что структура модели $\boldsymbol{\Gamma}$ для вероятностной модели глубокого обучения $\mathbf{f}$ и метапараметры $\boldsymbol{\lambda}$ определены однозначно (5).

Пусть априорное распределение параметров имеет вид (6). Требуется найти параметры $\boldsymbol{\theta}^*$ и гиперпараметры $\mathbf{h}^*$ модели, доставляющие максимум следующей функции:

$$\mathbf{h}^* = \arg\max_{\mathbf{h} \in \mathbb{H}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*, \boldsymbol{\lambda}), \quad \boldsymbol{\theta}^*(\mathbf{h}) = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}), \tag{14}$$

где L, Q — функции потерь и валидации.

Перечислим алгоритмы оптимизации гиперпараметров, исследованные в этой главе:

1. Случайный поиск, стохастический метод. Неэффективен при большом количестве гиперпараметров в силу проклятия размерности.
2. Жадный алгоритм (Luketina: 2016), градиентный метод. Доставляет локально-оптимальное решение задачи оптимизации. Позволяет производить одновременную оптимизацию параметров и гиперпараметров.
3. HOAG (Pedregosa: 2016), градиентный метод. Алгоритм основан на приближении аналитического решения двухуровневой задачи оптимизации (14).
4. DrMAD (Fu: 2017), градиентный метод. Алгоритм работает в строгих предположениях о линейности траектории оптимизации гиперпараметров. Алгоритм позволяет также производить оптимизацию параметров оператора оптимизации.

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. Показано, 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 ($\mathcal{GS}$): $p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) = \Pi_{(j,k)\in E}\, p(\boldsymbol{\gamma}^{j,k}|\mathbf{s}^{j,k}, \lambda_{\text{temp}})$, 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:

$$(K^{j,k}-1)!(\lambda_{\text{temp}})^{K^{h,j}-1} \prod_{l=0}^{K^{j,k}-1} s_l^{j,k} (\boldsymbol{\gamma}_l^{j,k})^{-\lambda_{\text{temp}}-1} \left(\sum_{l=0}^{K^{j,k}-1} s_l^{j,k} (\boldsymbol{\gamma}_l^{j,k})^{-\lambda_{\text{temp}}} \right)^{-K^{j,k}}, \quad (15)$$

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.

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 $\boldsymbol{\Gamma}$, содержащая все структурные параметры $\{\boldsymbol{\gamma}^{j,k}, (j,k) \in E\}$, распределенные по распределению Gumbel-Softmax.
3. Гиперпараметры $\mathbf{h} = [\text{diag}(\mathbf{A}), \mathbf{s}]$, 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:

$$p(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\lambda}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) p(\mathbf{h}|\boldsymbol{\lambda}) \to \max_{\mathbf{h} \in \mathbb{H}}. \tag{16}$$

Структура и параметры модели выбираются на основе полученных значений гиперпараметров: $\mathbf{w}^*, \boldsymbol{\Gamma}^* = \arg\max_{\mathbf{w} \in \mathbb{W}, \boldsymbol{\Gamma} \in \mathbb{\Gamma}} p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}^*, \boldsymbol{\lambda})$, где $\mathbf{h}^*$ — решение задачи оптимизации (16).

Для вычисления обоснованности модели $p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ из (16) предлагается использовать нижнюю вариационную оценку обоснованности.

Теорема 3. Пусть $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) = q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ — вариационное распределение с параметрами $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\mathbf{w}}, \boldsymbol{\theta}_{\boldsymbol{\Gamma}}]$, аппроксимирующее апостериорное распределение структуры и параметров:

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

Тогда справедлива следующая оценка:

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) \geq \tag{17}$$

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

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

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

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

Определение 16. 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 17. 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:

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

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

$$\rho(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda}) = \prod_{w\in\mathbf{w}} \rho(w|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda}). \tag{18}$$

Теорема 4. Пусть

1. Заданы компактные множества $U_{\mathbf{h}} \subset \mathbb{H}, U_{\boldsymbol{\theta}_{\mathbf{w}}} \subset \Theta_{\mathbf{w}}, U_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}} \subset \Theta_{\boldsymbol{\Gamma}}$.
2. Вариационное распределение $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ является абсолютно непрерывным и унимодальным на $U_{\boldsymbol{\theta}}$. Его мода и матожидание совпадают.
3. Априорное распределение $p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ является абсолютно непрерывным и унимодальным на $U_{\mathbf{h}}$. Его мода и матожидание совпадают и не зависят от гиперпараметров $\mathbf{h}$ на $U_{\mathbf{h}}$ и структуры $\boldsymbol{\Gamma}$ на $U_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}}$:

$$\mathsf{E}_{p(\mathbf{w}|\boldsymbol{\Gamma},\mathbf{h},\boldsymbol{\lambda})}\ \mathbf{w} = \text{mode } p(\mathbf{w}|\boldsymbol{\Gamma}_1, \mathbf{h}_1, \boldsymbol{\lambda}) = \text{mode } p(\mathbf{w}|\boldsymbol{\Gamma}_1, \mathbf{h}_2, \boldsymbol{\lambda}) = \mathbf{m}$$

   для любых $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2 \in U_{\boldsymbol{\Gamma}}$.
4. Параметры модели $\mathbf{w}$ имеют конечные вторые моменты по маргинальным распределениям: $\int_{\boldsymbol{\Gamma}} q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) d\boldsymbol{\Gamma}$, $\int_{\boldsymbol{\Gamma}} q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) d\boldsymbol{\Gamma}$ при любых $\boldsymbol{\theta}_{\mathbf{w}} \in U_{\boldsymbol{\theta}_{\mathbf{w}}}, \boldsymbol{\theta}_{\boldsymbol{\Gamma}} \in U_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}}, \mathbf{h} \in U_{\mathbf{h}}$.
5. Вариационное распределение $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ является липшицевым по $\mathbf{w}$.
6. Значение $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ не равно нулю при любых $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}, \boldsymbol{\Gamma} \in \mathbb{\Gamma}$.
7. Точная нижняя грань $\inf_{\boldsymbol{\Gamma}\in\mathbb{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}\in U_{\boldsymbol{\theta}_{\mathbf{w}}}} q_{\mathbf{w}}(\mathbf{m}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ не равна нулю.
8. Решение задачи $\mathbf{h}^* = \arg\min_{\mathbf{h}\in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big)$ единственно для любого $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$.

9. Задана бесконечная последовательность векторов вариационных параметров $\boldsymbol{\theta}[1], \boldsymbol{\theta}[2], \ldots, \boldsymbol{\theta}[i], \cdots \in U_{\boldsymbol{\theta}}$, такая, что $\lim_{i\to\infty} C_p(\boldsymbol{\theta}[i]|U_{\mathbf{h}}, \boldsymbol{\lambda}) = 0$.

Тогда следующее выражение стремится к единице: $\lim_{i\to\infty} \mathsf{E}_{q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i])} \rho(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}[i], \mathbf{h}[i], \boldsymbol{\lambda})^{-1}$.

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:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log\ p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}). \tag{19}$$

Метод не предполагает оптимизации гиперпараметров $\mathbf{h}$. Для формального соответствия данной задачи задаче выбора модели (4), 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:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \tag{20}$$

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

Т.к. в рамках данной задачи (20) 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:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) = p'|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) \tag{21}$$

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

4. Критерий Акаике: $\text{AIC} = 2\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - 2|\mathbb{W}| \to \max$. Для использования критерия Акаике для сравнения моделей, принадлежащих одному параметрическому семейству $\mathfrak{F}$ предлагается следующая переформулировка:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \tag{22}$$

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

где

$$\mathbf{h} = \arg\min_{\mathbf{h}' \in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big), \tag{23}$$

λ_{prune} — метапараметр алгоритма, $U_{\mathbf{h}} \subset \mathbb{H}$ — область определения задачи по гиперпараметрам. Предложенное обобщение (22) применимо только в случае, если выражение (23) 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:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \log \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \tag{24}$$

$$-0.5\log m|\{w : D_{\mathrm{KL}}\big(q_{\mathbf{w}}(w|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})||p(w|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})\big)\}|,$$

метапараметр λ_{prune} определен аналогично (23).

6. Метод вариационной оценки обоснованности:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \tag{25}$$

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

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

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:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}}, \mathbf{w}, \boldsymbol{\Gamma}) + \log p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) \to \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}}, \tag{26}$$

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

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 18. 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 (4) $\boldsymbol{\theta}^*(\mathbf{h}) = \arg\max_{\boldsymbol{\theta}\in\mathbb{O}} L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$ 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 оптимизации гиперпараметров, соответствующих критерию: $\mathbf{h} = [\mathbf{A},\mathbf{s}]$, 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 $K_3 > 0$ и набор метапараметров $\boldsymbol{\lambda}$, такие, 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 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 $\boldsymbol{\lambda}' \neq \boldsymbol{\lambda}$, такое, 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 $\boldsymbol{\lambda} \in U_{\boldsymbol{\lambda}}$.

Теорема 5. Рассмотренные задачи (19),(20),(21),(22),(24),(26) не являются обобщающими.

Теорема 6. Пусть $q_{\boldsymbol{\Gamma}}$ — абсолютно непрерывное распределение с дифференцируемой плотностью, такой, что:

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

Тогда задача (25) не является обобщающей.

В качестве обобщающей задачи оптимизации предлагается оптимизационную задачу следующего вида:

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \tag{27}$$

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

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \tag{28}$$

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

где $\mathfrak{P}$ — непустое множество распределений на структуре $\boldsymbol{\Gamma}$, $\lambda^{\mathrm{Q}}_{\mathrm{prior}}, \lambda^{\mathrm{L}}_{\mathrm{prior}}, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$ — некоторые числа. Множество распределений $\mathfrak{P}$ отвечает за перебор структур $\boldsymbol{\Gamma}$ 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.

Теорема 7. 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 $\mathbf{s} \in U_{\mathbf{h}}, \lambda_{\mathrm{temp}} \in U_{\boldsymbol{\lambda}}$, 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 $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}, \lambda^{\mathrm{Q}}_{\mathrm{prior}}, \lambda^{\mathrm{L}}_{\mathrm{prior}}, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$ включает отрезок от нуля до единицы.
7. Существует значение метапараметров $\lambda_1 > 0, \lambda_2 > 0, \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} > 0 \in U_{\boldsymbol{\lambda}}$, 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 $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}, \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 0$.
8. Существует значение метапараметров $\lambda^{\mathrm{L}}_{\mathrm{prior}} > 0, \lambda^{\mathrm{Q}}_{\mathrm{prior}} > 0, \lambda_1 > 0, \lambda_2 > 0, \lambda_{\mathrm{temp}} > 0 \in U_{\boldsymbol{\lambda}}$, такое, что

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

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

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

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

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

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

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 (27) 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 8. 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 (27) 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 9. 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 10. Пусть

1. Задан компакт $U = U_{\mathbf{h}} \times U_{\boldsymbol{\theta}}$ и $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$.

2. Решение задачи

$$\min_{\mathbf{h}\in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) \tag{29}$$

является единственным для некоторых $\lambda^{\mathrm{Q}}_{\mathrm{prior}_1}, \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}, \lambda^{\mathrm{Q}}_{\mathrm{prior}_1} > \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$ на U при некоторых фиксированных $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}, \lambda^{\mathrm{L}}_{\mathrm{prior}}, \lambda_{\mathrm{temp}}, \lambda_1, \lambda_2$.

3. Решения задачи (27),(28) являются единственными на U при $\lambda^{\mathrm{Q}}_{\mathrm{prior}_1}, \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$ и $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}, \lambda^{\mathrm{L}}_{\mathrm{prior}}, \lambda_{\mathrm{temp}}, \lambda_1, \lambda_2$.
4. Функция $Q(\mathbf{h}|\boldsymbol{\theta}_2, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ является вогнутой по $\mathbf{h} \in U_{\mathbf{h}}$ при $\lambda^{\mathrm{Q}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$.
5. Решение задачи (29) единственно при $\lambda^{\mathrm{Q}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$.
6. 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.
7. Значения $p(\mathbf{h}|\boldsymbol{\lambda})$ приблизительно равны на $U_{\mathbf{h}}$: $p(\mathbf{h}_1|\boldsymbol{\lambda}) \approx p(\mathbf{h}_2|\boldsymbol{\lambda})$ для всех $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}$.

Тогда справедлива следующая оценка разности параметрических сложностей:

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

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

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

Теорема 11. Пусть $\frac{\lambda^{\mathrm{Q}}_{\mathrm{prior}}}{\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}} = \lambda^{\mathrm{L}}_{\mathrm{prior}}$. Тогда задача оптимизации (27) 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, задаче 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.
7. Рассмотрено применение предложенных методов для построения моделей глубокого обучения в прикладных задачах регрессии и классификации.

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.

Прочие 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.