Московский физико-технический институт
(государственный университет)

На правах рукописи

Бахтеев Олег Юрьевич

# Последовательный выбор моделей глубокого обучения оптимальной сложности

05.13.17 — Теоретические основы информатики

Диссертация на соискание ученой степени
кандидата физико-математических наук

Научный руководитель:
д.ф.-м.н. В. В. Стрижов

Москва — 2019

## Оглавление

## Введение

**Актуальность темы.** В работе рассматривается задача автоматического построения моделей глубокого обучения субоптимальной сложности.

Под сложностью модели понимается *минимальная длина описания* [1], 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* [1]. Для получения оценки обоснованности используются вариационные методы получения оценки обоснованности [2], основанные на аппроксимации неизвестного другим заданным распределением. Под субоптимальной сложностью понимается вариационная оценка обоснованности модели.

Одна из проблем построения моделей глубокого обучения — большое количество параметров моделей [3, 4]. Поэтому задача выбора моделей глубокого обучения включает в себя выбор стратегии построения модели, эффективной по вычислительным ресурсам. В работе [5] приводятся теоретические оценки построения нейросетей с использованием , при которых построение модели производится итеративно последовательным увеличением числа нейронов в сети. В работе [6] предлагается жадная стратегия выбора модели нейросети с использованием релевантных априорных распределений, т.е. параметрических распределений, оптимизация параметров которых позволяет удалить часть параметров из модели. Данный метод был к задаче построения модели метода релевантных векторов [7]. Альтернативой данным алгоритмам построения моделей являются методы, основанные на прореживании сетей глубокого обучения [8, 9, 10], т.е. последовательного удаления параметров, не дающих существенного прироста качества модели. В работах [11, 12] рассматривается послойное построение модели с отдельным критерием оптимизации для каждого слоя. В работах [13, 14, 15] предлагается декомпозиция модели на порождающую и разделяющую, оптимизируемых последовательно. В работе [16] 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 [17, 18, 19, 20] предлагается метод автоматического построения сверточной сети с использованием обучения с подкреплением. В [21] используется схожее представление сверточной сети, вместо обучения с подкреплением используется градиентная параметров, задающих структуру нейронной сети.

В качестве порождающих моделей в сетях глубокого обучения выступают ограниченные машины Больцмана [3] и автокодировщики [22]. В работе [23] рассматриваются некоторые типы регуляризации автокодировщиков, позволяющие формально рассматривать данные модели как порождающие модели с использованием байесового вывода. В работе [24] также рассматриваются ре-

гуляризованные автокодировщики и свойства оценок их правдоподобия. В работе [25] предлагается обобщение автокодировщика с использованием вариационного байесовского вывода [2]. В работе [26] рассматриваются модификации вариационного автокодировщика и ступенчатых сетей (англ. ladder network) [27] для случая построения многослойных порождающих моделей.

В качестве критерия выбора модели в ряде работ [28, 2, 29, 30, 31, 32] выступает обоснованность модели. В работах [29, 30, 31, 32] рассматривается проблема выбора модели и оценки гиперапараметров в задачах регрессии. Альтернативным критерием выбора модели является минимальная длина описания [1], являющаяся показателем статистической сложности модели и заданной выборки. В работе [33] рассматривается перечень критериев сложности моделей глубокого обучения и их взаимосвязь. В работе [34] в качестве критерия сложности модели выступает показатель нелинейности, характеризуемый степенью полинома Чебышева, аппроксимирующего функцию. В работе [35] 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 [36] 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 [37], в которой исследуется устойчивость классификации модели под действием шума.

Одним из методов получения приближенного значения обоснованности является вариационный метод получения нижней оценки интеграла [2]. В работе [38] рассматривается стохастическая версия вариационного метода. В работе [39] рассматривается алгоритм получения вариационной нижней оценки обоснованности для оптимизации гиперпараметров моделей глубокого обучения. В работе [40] рассматривается получение вариационной нижней оценки интеграла с использованием модификации методов Монте-Карло. В работе [41] рассматривается стохастический градиентный спуск в качестве оператора, порождающего распределение, аппроксимирующее апостериорное распределение параметров модели. Схожий подход рассматривается в работе [42], где также рассматривается стохастический градиентный спуск в качестве оператора, порождающего апостериорное распределение параметров. В работе [43] предлагается модификация стохастического градиентного спуска, аппроксимирующая апостериорное распределение.

Альтернативным методом выбора модели является выбор модели на основе скользящего контроля [44, 29]. Проблемой такого подхода является возможная высокая вычислительная сложность [45, 46]. В работах [47, 48] рассматривается проблема смещения оценок качества модели и гиперпараметров, получаемых при использовании $k$-fold метода скользящего контроля, при котором выборка делится на $k$ частей с обучением на $k-1$ части и валидацией результата на оставшейся части выборки.

Задачей, связанной с проблемой выбора модели, является задача оптимизации гиперпараметров [28, 2]. В работе [29] рассматривается оптимизация гиперпарамтров с использованием метода скользящего контроля и методов оптимизации обоснованности моделей, отмечается низкая скорость сходимости гиперпараметров при использовании метода скользящего контроля. В ряде работ [49, 50] рассматриваются градиентные методы оптимизации гиперпараметров, позволяющие оптимизировать большое количество гиперпараметров одновременно. В работе [49] предлагается метод оптимизации гиперпараметров с использованием градиентного спуска с моментом, в качестве оптимизируемого функционала рассматривается ошибка на валидационной части выборки. В работе [51] предлагается метод аппроксимации градиента функции потерь по гиперпараметрам, позволяющий использовать градиентные методы в задаче оптимизации гиперпараметров на больших выборках. В работе [52] предлагается упрощенный метод оптимизации гиперпараметров с градиентным спуском: вместо всей истории обновлений параметров для оптимизации используется только последнее обновление. В работе [42] рассматривается задача оптимизации параметров градиентного спуска с использованием нижней вариационной оценки обоснованности.

**Цели работы.TODO**

1. Исследовать методы построения моделей глубокого обучения оптимальной сложности.
2. Предложить критерии оптимальной и субоптимальной сложности модели глубокого обучения.
3. Предложить метод выбора структуры модели глубокого обучения.
4. Предложить алгоритм построения модели субоптимальной сложности и оптимизации параметров.
5. Разработать алгоритм построения модели и проанализировать различные подходы к решению задачи автоматического построения моделей глубокого обучения и оптимизации параметров модели.

**Методы исследования.** Для достижения поставленных целей используются методы вариационного байесовского вывода [28, 2, 41]. Рассматриваются графовое представление нейронной сети [17, 21]. Для получения вариационных оценок обоснованности модели используется метод, основанный на градиентном спуске [42, 41]. В качестве метода получения модели субоптимальной сложности используется метод автоматического определения релевантности параметров [28, 53] с использоваением градиентных методов оптимизации гиперпараметров [49, 50, 52, 51].

**Основные положения, выносимые на защиту.**

1. Предложен метод байесовского выбора субоптимальной структуры модели глубокого обучения с использованием автоматического определения релевантности параметров.
2. Предложен метод графового описания моделей глубокого обучения.
3. Проведено исследование свойства оптимизационных алгоритмов выбора модели.
4. Предложена обобщающая задача оптимизации модели, обобщающая 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 на основе мультистарта оптимизации модели
6. TODO Предложен алгоритм оптимизации параметров, гиперпараметров и структурных параметров моделей глубокого обучения.
7. Проведено исследование свойств оптимизационной задачи при различных значениях метапараметров. Рассмотрены ее ассимптотические свойства.

**Научная новизна.** Разработан новый подход к построению моделей глубого обучения. Предложены критерии субоптимльной и оптимальной сложности модели, а также исследована их связь. Предложен метод построения модели глубокого обучения субоптимальной сложности. TODO Предложен метод оптимизации гиперпараметров модели, а также методов оптимизации 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.

TODO: p(...|f)

**Практическая значимость.** Предложенные в работе методы предназначены для построения моделей глубокого обучения в задачах регрессии и классификации; оптимизации гиперпараметров полученной модели; выборе 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 конеренция «57-я научная конеренция МФТИ», 2014.
2. "A monolingual approach to detection of text reuse in Russian-English collection", Международная конференция «Artificial Intelligence and Natural Language Conference», 2015 [54].
3. "Выбор модели глубокого обучения субоптимальной сложности с использованием вариационной оценки правдоподобия", Международная конференция «Интеллектуализация обработки информации», 2016 [55].
4. "Machine-Translated Text Detection in a Collection of Russian Scientific Papers", Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог-21», 2017 [56].
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 [57].
7. "Детектирование переводных заимствований в текстах научных статей из журналов, входящих в РИНЦ", Всероссийская конференция «Математические методы распознавания образов ММРО», 2017 [58].
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 [59].
10. "Variational learning across domains with triplet information", Международная конференция «Visually Grounded Interaction and Language workshop, Conference on Neural Information Processing Systems», 2018.

Работа поддержана грантами Российского фонда фундаментальных исследований.

1. 19-07-00875, Развитие методов автоматического построения и выбора вероятностных моделей субоптимальной сложности в задачах глубокого обучения.
2. 16-37-00488, Разработка алгоритмов построения сетей глубокого обучения как суперпозиций универсальных моделей.

3. 16-07-01158, Развитие теории построения суперпозиций универсальных моделей классификации сигналов.
4. 14-07-3104, Построение и анализ моделей классификации для выборок малой мощности.

**Публикации по теме диссертации.** Основные результаты по теме диссертации изложены в 8 печатных изданиях, 6 из которых изданы в журналах, рекомендованных ВАК.

1. Бахтеев О.Ю., Попова М.С., Стрижов В.В. Системы и средства глубокого обучения в задачах классификации. // Системы и средства информатики. 2016. № 26.2. С. 4-22 [9].
2. Бахтеев О.Ю., Стрижов В.В. Выбор моделей глубокого обучения субоптимальной сложности. // Автоматика и телемеханика. 2018. №8. С. 129-147 [60].
3. Огальцов А.В., Бахтеев О.Ю. Автоматическое извлечение метаданных из научных PDF-документов. // Информатика и её применения. 2018 [**?**].
4. Смердов А.Н., Бахтеев О.Ю., Стрижов В.В. Выбор оптимальной модели рекуррентной сети в задачах поиска парафраза. // Информатика и ее применения. 2019.
5. Грабовой А.В., Бахтеев О.Ю., Стрижов В.В. Определение релевантности параметров нейросети. // Информатика и её применения. 2019.
6. Bakhteev O., Strijov V. Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Annals of Operations Research. 2019 TODO [61].
7. 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 [62].
8. Бахтеев О.Ю. Восстановление пропущенных значений в разнородных шкалах с большим числом пропусков. // Машинное обучение и анализ данных. 2015. № 11. С. 1-11 [63].

**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 125 наименований. Основной текст занимает 88 страниц.

**Краткое содержание работы по главам.** В первой главе вводятся основные понятия и определения, формулируются задачи построения моделей глубокого обучения. Рассматриваются основные критерии выбора моделей. Рассматриваются существующие алгоритмы построения моделей глубокого обучения.

Во второй главе предлагается алгоритм построения суботпимальной модели глубокого обучения. Предлагаются методы оценки сложности модели.

В третьей главе рассматриваются методы оптимизации гиперпараметров модели.

В четвертой главе рассматривается обобщенный метод выбора модели глубокого обучения.

TODO В пятой главе на базе предложенных методов описывается разработанный программный комплекс, позволяющий автоматически построить модель глубокого обучения субпотимальной сложности для заданной выборки для задачи классификации и регрессии. Работа данного комплекса анализируется на ряде выборок для задач классификации и регрессии. Результаты, полученные с помощью предложенных методов, сравниваются с результатами известных алгоритмов.

# Глава 1
# Постановка задачи последовательного выбора моделей

Проблема выбора структуры модели является фундаментальной в области машинного обучения интеллектуального анализа данных. 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

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i = 1, \ldots, m, \tag{1.1}$$

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:

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

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

Специфика задачи выбора модели *глубокого обучения* заключатеся 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

$$\mathbf{f}_{v_k}(\mathbf{w}, \mathbf{x}) = \mathbf{agg}_{v_k}\left(\{\langle \boldsymbol{\gamma}^{j,k}, \mathbf{g}^{j,k}\rangle \circ \mathbf{f}_j(\mathbf{x}) | j \in \mathrm{Adj}(v_k)\}\right), \tag{1.2}$$

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

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 множестом векторов базовых функций $\{\mathbf{g}^{j,k}, (j,k) \in E\}$ и функций агрегаций $\{\mathbf{agg}_v, v \in V\}$ 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 (1.2) 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.

*Доказательство.* Утверждение следует непосредственно из определения: по условию утверждения для любого $\boldsymbol{\gamma}^{j,k} \in [0,1]^{K^{j,k}}$ функция является дифференцируемой функцией из признакового пространства $\mathbb{X}$ в пространство меток $\mathbb{Y}$, что соответствует определению модели. □

Пример параметрического семейства моделей, которое описывает сверточную нейронную сеть, представлена на Рис. 1.1. Семейство задает множество моделей с двумя операциями свертки с одинаковым размером фильтра $c_0$ и различным числом каналов $c_1$ и $c_2$. Единичная свертка с $c_1$ каналами $\mathbf{Conv}(\mathbf{x}, c_1, 1)$ требуется для выравнивания размерностей скрытых слоев. Каждая модель параметрического семейства задается формулой:

$$\mathbf{f} = \mathbf{agg}_2\left(\left\{\gamma_0^{1,2}\mathbf{g}_0^{1,2}\left(\mathbf{agg}_1\left(\{\gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}), \gamma_1^{0,1}\mathbf{g}_1^{0,1}(\mathbf{x})\}\right)\right)\right\}\right).$$

Положим, что функции агрегации $\mathbf{agg}_1, \mathbf{agg}_2$ являются операциями суммы. Заметим, что к вершине "2" ведет только одно ребро, поэтому операцию суммы можно опустить. Итоговая формула модели задается следующим образом:

$$\mathbf{f} = \gamma_0^{1,2}\mathbf{softmax}\big(\gamma_0^{0,1}\mathbf{Conv}(\mathbf{x}, c_0, c_1)(\mathbf{x}) + \\ + \gamma_1^{0,1}\mathbf{Conv}(\mathbf{x}, 1, c_1) \circ \mathbf{Conv}(\mathbf{x}, c_0, c_2)(\mathbf{x})\mathbf{w}_0^{1,2}\big).$$

**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 $\mathbf{w}_l^{j,k}$.

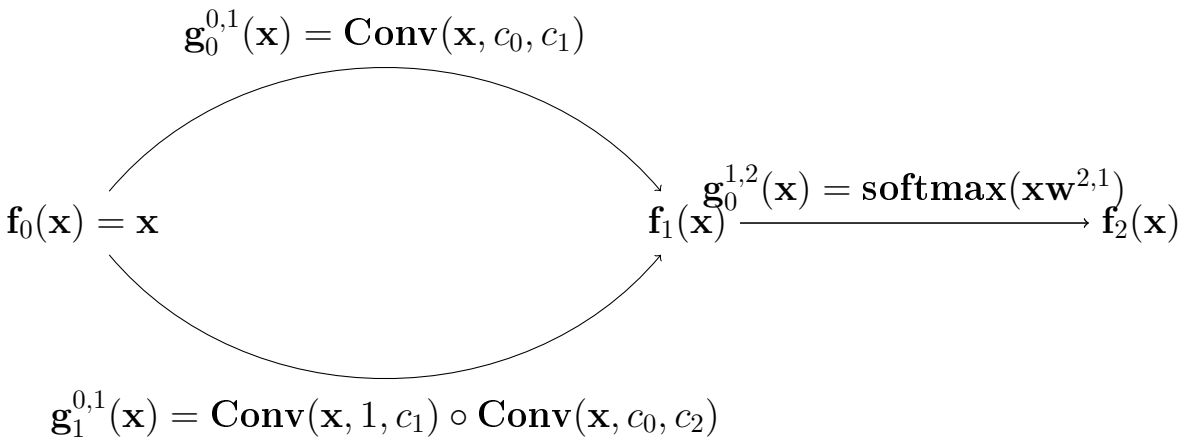



Рис. 1.1. Пример параметрического семейства моделей глубокого обучения: семейство описывает сверточную нейронную сеть.

**Определение 6.** Структурой $\mathbf{\Gamma}$ модели $\mathbf{f}$ из параметрического семейства моделей $\mathfrak{F}$ назовем конкатенацию векторов $\boldsymbol{\gamma}^{j,k}$. Множество всех возможных значений структуры $\mathbf{\Gamma}$ 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 $\mathbf{\Gamma}$ при котором функция $\mathbf{f}$ совпадает с функцией (1.2).

Предложенное определение параметризации не противоречит определению параметризации глубоких моделей в других работах. В [35] под параметризацией понимается представление матрицы параметров модели с использованием аппроксимации низкоранговыми матрицами. В [64] под параметризацией модели глубокого обучения понимается выбор графа, позволяющего описать структуру заданной модели глубокого обучения.

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

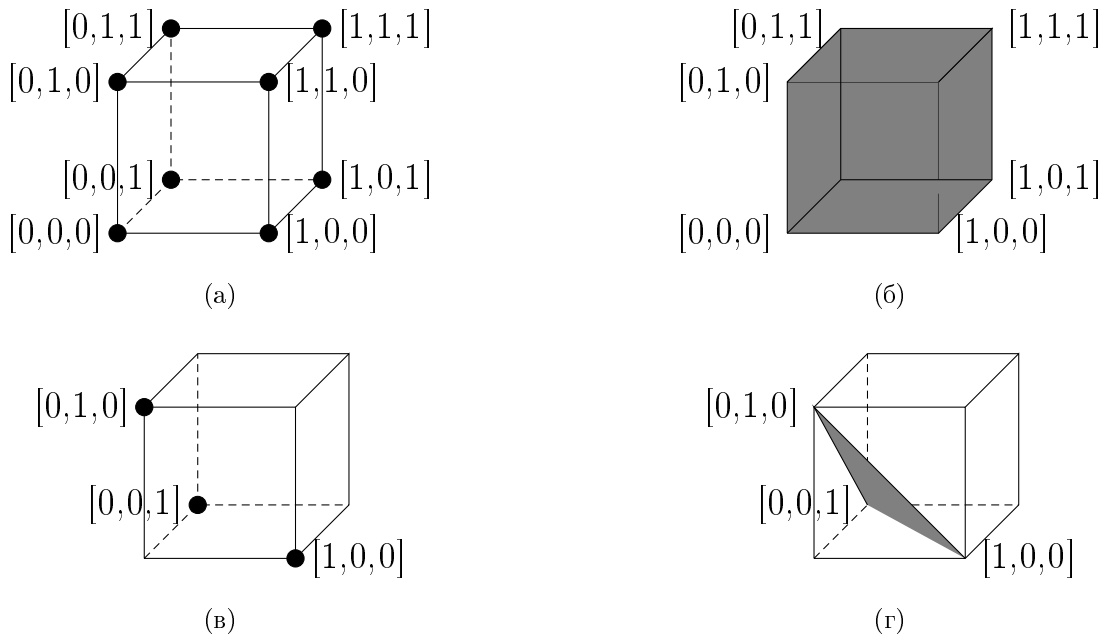



Рис. 1.2. Примеры ограничений для одного структурного параметра $\boldsymbol{\gamma}^{j,k}$, $K^{j,k} = 3$.
а) структурный параметр лежит на вершинах куба, б) структурный параметр лежит внутри куба, в) структурный параметр лежит на вершинах симплекса, г) структурный параметр лежит внутри симплекса.

итоговую модель. Примером параметрического семейства моделей, требующим такое ограничение является семейство полносвязанных нейронных сетей с одним скрытым слоем и двумя значениями коилчества нейронов на скрытом слое. Схема семейства представлена на Рис. 1.5. Данное семейство можно представить как семейство с двумя базовыми функциями вида $\mathbf{g} = \boldsymbol{\sigma}(\mathbf{w}^{\intercal}\mathbf{x})$, 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.

Пример, иллюстрирующий представленные выше ограничения, изображен на Рис. 1.2. 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 [21].

Для дальнейшей постановки задачи введем понятие вероятностной модели,

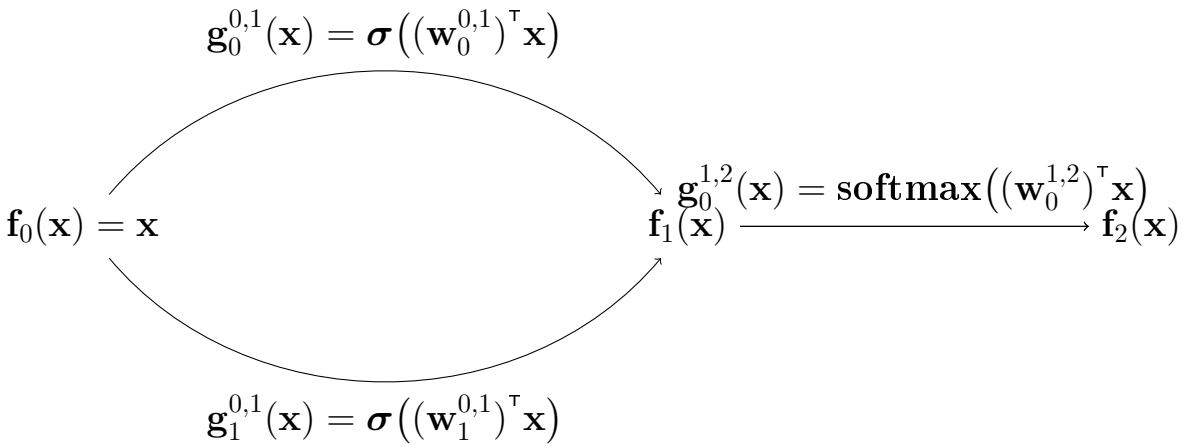



Рис. 1.3. Пример параметрического семейства моделей глубокого обучения: семейство описывает многослойную полносвязную нейронную сеть с одним скрытым слоем и нелинейной функцией активации $\boldsymbol{\sigma}$.

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:

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

где $\mathbb{W}$ — множество значений параметров модели, $\mathbb{H}$ — множество значений гиперпараметров, $\mathbb{\Lambda}$ – множество значений метапараметров. Формальное определение последних 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

$$p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})}{p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})} \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}). \tag{1.3}$$

**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

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

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

**Определение 13.** *Обоснованностью модели* назовем величину

$$p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \iint_{\mathbf{w},\mathbf{\Gamma}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) d\mathbf{w} d\mathbf{\Gamma}. \tag{1.4}$$

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 [29] и вариационная нижняя оценка [39]. В данной работе в качестве метода получения оценок обоснованности модели выступает вариационная нижняя оценка.

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

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

**Определение 16.** 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 $q$.

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

**Определение 17.** Пусть задано вариационное распределения $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})$ 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.

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

$$\mathbf{h}^* = \arg\min_{\mathbf{h} \in \mathbb{H}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^*, \boldsymbol{\lambda}), \tag{1.5}$$

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

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}). \tag{1.6}$$

**Определение 18.** *Задачей выбора модели* $\mathbf{f}$ назовем двухуровневую задачу оптимизации (1.5),(1.6).

Рассмотрим для примера базовый вариант выбора модели с применением функций $q, L, Q$.

**Пример 1.** Положим, что задано разбиение выборки на обучающую $\mathfrak{D}_{\text{train}}$ и валидационную $\mathfrak{D}_{\text{valid}}$ части. Положим в качестве вариационных параметров $\boldsymbol{\theta}$ параметры $\mathbf{w}$ и структуры $\boldsymbol{\Gamma}$ модели:

$$\boldsymbol{\theta} = [\mathbf{w}, \boldsymbol{\Gamma}].$$

Пусть также задано априорное распределение $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$. Положим в качестве функции $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ минус логарифм апостериорной вероятности модели:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = -\sum_{\mathbf{x}, y \in \mathfrak{D}_{\text{train}}} \log p(y, \mathbf{w}, \boldsymbol{\Gamma}|\mathbf{x}, \boldsymbol{\lambda}).$$

Положим в качестве функции $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ минус логарифм правдоподобия выборки при условии параметров $\mathbf{w}$ и структуры $\boldsymbol{\Gamma}$:

$$Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = -\sum_{\mathbf{x}, y \in \mathfrak{D}_{\text{valid}}} \log\ p(y|\mathbf{x}, \mathbf{w}, \boldsymbol{\Gamma}, \boldsymbol{\lambda}).$$

Оптимизация параметров и структуры производится по обучающей выборке. Гиперпараметры $\mathbf{h}$ выступают в качестве регуляризатора, чья оптимизация производится по валидационной выборке. Подобная оптимизация позволяет предотвратить переобучение модели [49].

Частным случаем задачи выбора структуры глубокой сети является выбор обобщенно-линейных моделей. Отдельные слои полносвязанных нейросетей являются обобщенно-линейными модели. Задачу выбора обобщенно-линейной моделей сводится к задаче выбора признаков, методы решения которой делятся на три группы [65]:

1. Фильтрационные методы. Не используют какой-либо информации о модели, а отсекают признаки только на основе статистических показателей, учитывающих взаимосвязь признаков и меток объектов.
2. Оберточные методы анализируют подмножества признаков. Они выбирают не признаки, а подмножества признаков, что позволяет учесть корреляция признаков.
3. Методы погружения оптимизируют модели и проводят выбор признаков в единой процедуре, являясь комбинацией предыдущих типов отбора признаков.

## 1.1. Критерии выбора модели глубокого обучения

В данном разделе рассматриваются различные критерии выбора моделей глубокого обучения, соответствующие функции валидации $Q$. В данной работе в качестве критерия выбора модели предлагается субоптимальная сложность

модели. Под сложностью модели понимается *обоснованность модели* (1.4), являющееся байесовской интерпретацией *минимальной длины описания* [1], т.е. минимальное количество информации, которое требуется передать о модели и о выборке:

$$\mathrm{MDL}(\mathbf{y}, \mathbf{f}) = \mathrm{Len}(\mathbf{y}|\mathbf{w}^*, \mathbf{f}) + \mathrm{COMP}(\mathbf{f}), \tag{1.7}$$

где $\mathrm{Len}(\mathbf{y}|\mathbf{w}^*, \mathbf{f})$ — *длина описания* матрицы $\mathbf{y}$ с использованием модели $\mathbf{f}$ и оценки вектора параметров $\mathbf{w}^*$, полученных методом наибольшего правдоподобия, а $\mathrm{COMP}(\mathbf{f})$ — величина, характеризующая *параметрическую сложность* модели, т.е. способность модели описать произвольную выборки из $\mathbb{X}$ [1].

В общем случае правдоподобие модели является трудновычислимым. Для получения оценки правдоподобия используются вариационные методы получения оценки правдоподобия [2], основанные на аппроксимации неизвестного другим заданным распределением. Под субоптимальной сложностью понимается вариационная оценка правдоподобия модели. Альтернативной величиной, характеризующей сложность модели, выступает рademахеровская сложность (1.14). Данная величина используется как критерий для продолжения итеративного построения модели в [16].

В работе [33] рассматривается ряд критериев сложности моделей глубокого обучения и их взаимосвязь. В работе [34] в качестве критерия сложности модели выступает показатель нелинейности, характеризуемый степенью полинома Чебышева, аппроксимирующего функцию. В работе [35] анализируется показатель избыточности параметров сети. Утверждается, что по небольшому набору параметров в глубокой сети с большим количеством избыточных параметров возможно спрогнозировать значения остальных. В работе [36] 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 [37], в которой исследуется устойчивость классификации модели под действием шума. В ряде работ [28, 2, 29, 30, 31, 32] в качестве критерия выбора модели выступает правдоподобие модели. В работах [29, 30, 31, 32] рассматривается проблема выбора модели и оценки гиперапараметров в задачах регрессии. Альтернативным критерием выбора модели является минимальная длина описания [1], являющаяся показателем статистической сложности модели и заданной выборки. В работе [1] рассматриваются различные модификации и интерпретации минимальной длины описания, в том числе связь с правдоподобием модели.

Одним из методов получения приближенного значения правдоподобия модели является вариационный метод получения нижней оценки правдоподобия [2]. В работе [38] рассматривается стохастическая версия вариационного метода. В [39] рассматривается алгоритм получения вариационной нижней оценки правдоподобия для оптимизации гиперпараметров моделей глубокого обучения. В работе [40] рассматривается взаимосвязь градиентных методов получения ва-

риационной нижней оценки интеграла с методом Монте-Карло. В [41] рассматривается стохастический градиентный спуск в качестве оператора, порождающего распределение, аппроксимирующее апостериорное распределение параметров модели. В работе отмечается, что стохастический градиентный спуск не оптимизирует вариационную оценку правдоподобия, а приближает ее только до некоторого числа итераций оптимизации. Схожий подход рассматривается в работе [42], где также рассматривается стохастический градиентный спуск в качестве оператора, порождающего апостериорное распределение параметров. В работе [43] предлагается модификация стохастического градиентного спуска, аппроксимирующая апостериорное распределение.

Альтернативным методом выбора модели является выбор модели на основе скользящего контроля [44, 29]. Проблемой такого подхода является высокая вычислительная сложность [45, 46]. В работах [47, 48] рассматривается проблема смещения оценок качества модели и гиперпараметров, получаемых при использовании $k$-fold метода скользящего контроля, при котором выборка делится на $k$-частей с обучением на $k-1$ части и валидацией результата на оставшейся части выборки.

## 1.2. Оптимизация параметров в задаче выбора структуры модели

Один из подходов к выбору оптимальной модели заключается в итеративном удалении наименее информативных параметров модели. В данном разделе собраны методы оптимизации структуры существующей модели.

**Алгоритмы прореживания параметров модели.** В [8] предлагается удалять неинформативные параметры модели. Для этого находится точка оптимума $\boldsymbol{\theta}^*$ функции $L$, и производится разложение функции $L$ в ряд Тейлора в окрестности $\boldsymbol{\theta}^*$:

$$L(\boldsymbol{\theta}^* + \Delta\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) - L(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \frac{1}{2}\Delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{H}\Delta\boldsymbol{\theta} + o(||\Delta\boldsymbol{\theta}||^3), \qquad (1.8)$$

где $\mathbf{H}$ — гессиан функции $L$. Связь между параметрами не учитывается, поэтому гессиан матрицы $L$ является диагональным. Положим в качестве операции удаления параметра замену его значения на ноль. Выбор наиболее неинформативного параметра сводится к задаче условной минимизации (1.8) при условиях вида

$$\theta_i + \Delta\theta_i = 0, \quad \theta_i \in \boldsymbol{\theta}.$$

В результате решения данной задачи минимизации каждому параметру определяется функция выпуклости

$$\text{saliency}(\theta_i) = \frac{\theta_i^2}{2(H^{-1})_{i,i}}.$$

Данная функция характеризует информативность параметра.

В [66] было предложено развитие данного метода. В отличие от [8] не вводится предположений о диагональности гессиана функции ошибок, поэтому удаление неинформативных параметров модели производится точнее. Для получения оценок гессиана и его обратной матрицы применяется итеративный алгоритм.

**Алгоритмы компрессии параметров модели.** В [67, 68, 10] предлагаются методы компрессии параметров сетей глубокого обучения. Основным отличием задачи прореживания от задачи компрессии выступает эксплуатационное требование: если прореживание используется для получения оптимальной и наиболее устойчивой модели, то компрессия производится для уменьшения потребляемых вычислительных ресурсов при сохранении основных эксплуатационных характеристик исходной модели [68]. В [10] предлагается итеритавиное использование регуляризации типа DropOut [69] для прореживания модели. В [67, 68] используются методы снижения вычислительной точности представления парамеров модели на основе кластеризации параметров $\mathbf{w}$ модели: вместо значений параметров предлагается хранить идентификатор кластера, соответствующего параметру, что существенно снижает количество требуемой памяти. В [68] предлагается метод компрессии, основанный на кластеризации значений параметров модели и представлении их в сжатом виде на основе кодов Хаффмана.

**Байесовские методы прореживания параметров модели.** Байесовский подход к порождению и выбору моделей заключается в использовании вероятностных предположений о распределении параметров и структуры в параметрических семействах моделей. Такой подход позволяет учитывать при выборе моделей не только эксплуатационные критерии качества модели, такие как точность итоговой модели и количество параметров в ней, но и некоторые статистические характеристики модели.

В работе [49] рассматривается задача оптимизации гиперпараметров. Авторы предлагают оптимизировать константы $l_2$-регуляризации отдельно для каждого параметра модели, проводится параллель с методами автоматического определения релевантности параметров (англ. automatic relevance determination, ARD) [28]. Идея автоматического определения релевантности заключается в выборе оптимальных начений гиперпараметров $\mathbf{h}$ с дальнейшим удалением неинформативных параметров. Неинформативными параметрами являются те параметры, которые с высокой вероятностью равны нулю относительно априорного или апостериорного распределения.

В работе [39] был предложен метод, основанный на получении вариационной нижней оценки правдоподобия модели. В качестве критерия информативности параметра выступает отношение вероятности нахождения параметра в пределах апостериорного распределения к вероятности равенства параметра нулю:

$$\rho = \exp\left(-\frac{\mu_j^2}{2\sigma_j^2}\right), \tag{1.9}$$

где $\mu_j, \sigma_j$ — среднее и дисперсия аппроксимирующего распределения $q$ для па-

раметра $w_j$.

Идея данного метода была развита в [70], где также используются вариационные методы. В отличие от [39], в [70] рассматривается ряд априорных распределений параметров, позволяющих прореживать модели более эффективно:

1. Нормальное распределение с лог-равномерным распределением дисперсии. Для каждого параметра $w \in \mathbf{w}$ задается группа параметров $\boldsymbol{\omega} \in \boldsymbol{\Omega}$, где $\boldsymbol{\Omega}$ — множество всех групп параметров:

$$p(\mathbf{w}|\mathbf{h}) \propto \prod_{\boldsymbol{\omega}_i \in \boldsymbol{\Omega}} \frac{1}{|\mathbf{h}_i|} \prod_{w \in \boldsymbol{\omega}_i} \mathcal{N}(w|\mathbf{0}, \mathbf{h}_i{}^2),$$

   где $\mathbf{h}_i$ — гиперпараметр, соответствующий группе $\omega_i$.
2. Априорное распределение задается произведением двух случайных величин $s_{\text{general}}, s_{jk}$ с половинным распределением Коши $\mathcal{C}^+$: одно ответственно за отдельный параметр, другое — за общее распределение параметров:

$$s_{\text{general}} \sim \mathcal{C}^+(0, h), \quad s_{jk} \sim \mathcal{C}^+(0, 1), \quad \hat{w}_{jk} \sim \mathcal{N}(0, 1), \quad w_{jk} \sim \hat{w}_{jk} s_{jk} s_{\text{general}},$$

   где $h \in \mathbf{h}$ — гиперпараметрв.

## 1.3. Оптимизация гиперпараметров модели

В данном разделе рассматриваются работы, посвященные методам оптимизации гиперпараметров. Методы, используемые для оптимизации гиперпараметров моделей глубокого обучения должны быть эффектинвыми по вычислительным затратам в силу высокой вычислительной сложности оптимизации параметров модели. В [71, 72] рассматривается задача оптимизации гиперпараметров стохастическими методами. В [71] проводится сравнение случайного поиска значений гиперпараметров с переборным алгоритмом. В [72] 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 $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}' = T(\boldsymbol{\theta}|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}), \tag{1.10}$$

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 [41].

Частным случаем оператора оптимизации является оператор стохастического спуска:

$$T(\boldsymbol{\theta}|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \boldsymbol{\theta} - \lambda_{\text{lr}} \nabla L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}), \tag{1.11}$$

где $\lambda_{\text{lr}}$ — 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 $\mathbf{h}$:

$$\mathbf{h} = T(\mathbf{h}|Q, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}). \tag{1.12}$$

В случае, если для решения задачи (1.6) применяется несколько шагов оператора оптимизации (1.10), $\boldsymbol{\theta}^*$ рассматривается как рекурсивная функция от начального приближения вариационных параметров $\boldsymbol{\theta}^0$ и вектора гиперпараметров $\mathbf{h}$:

$$\boldsymbol{\theta}^* = T \circ \cdots \circ T(\boldsymbol{\theta}|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \boldsymbol{\theta}^*(\boldsymbol{\theta}^0, \mathbf{h}). \tag{1.13}$$

Решение задачи оптимизации (1.12) при (1.13) является вычислительно сложным, поэтому применяются методы, аппроксимирующие применение градиентных методов при (1.13).

В [73] рассматривается оптимизация гиперпараметров градиентными методами для квадратичной функции потерь. В [49] в качестве оператора оптимизации гиперпараметров выступает метод градиентного спуска с моментом. Показано, что использование момента значительно снижает количество вычислительных ресурсов, требуемых для проведения оптимизации. В [74] предлагается аппроксимация градиентного метода, использующая предположение о линейности функции (1.13) от начального приближения $\boldsymbol{\theta}^0$. В [75] предлагается использовать численные методы для приближенного вычисления оператора оптимизации гиперпараметров. В [52] в качестве аппроксимации (1.13) предлагается рассматривать только последний шаг оптимизации:

$$\boldsymbol{\theta}^* \approx T(\boldsymbol{\theta}^{\eta-1}|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}),$$

где $\eta$ — число шагов оптимизации.

**Суррогатный выбор моделей.** Идея суррогатных моделей заключается в аппроксимации модели или параметрического семейства моделей вычислительно менее сложной функцией.

В работе [76] предлагается моделировать качество модели $Q$ (1.4) гауссовым процессом, параметрами которого выступают гиперпараметры исходной модели.

Одна из основных проблем использования гауссового процесса как суррогатной модели — кубическая сложность оптимизации. В работе [77] предлагается использовать случайные подпространства гиперпараметров для ускоренной оптимизации. В работе [78] предлагается кобминация из множества гауссовых моделей и линейной модели, позволяющая модели нелинейные зависимости гиперпараметров, а также существенно сократить сложность оптимизации.

В работе [79] предлагается рассматривать RBF-модель для аппроксимации качества $Q$ исходной модели, что позволяет ускорить процесс оптимизации суррогатной модели. В [80] рассматривается глубокая нейронная сеть в качестве

суррогатной функции. Вместо интеграла правдоподобия (1.4), который оценивается в случае использования гауссового процесса в качестве суррогата, используется максимум апостериорной веротяности (1.3).

Одним из параметров гауссовых процессов является функиия ядра гауссового процесса, полностью определяющая процесс в случае нулевого среднего. В работе [81] предлагается функция ядра, определенная на графах:

$$k(v_1, v_2) = r(d(v_1, v_2)),$$

где $d$ — геодезическое расстояние между вершинами графа, $r$ — некоторая вещественная функция, $v_1, v_2 \in V$.

В работе [82] рассматривается задач выбора структуры нейросети. Предлагается метод построения ковариационной функции для сравнения разнородных графов, соответствующих разным моделям нейронных сетей. Ковариационная функция основывается на метрике, заданной на некоторых числовых характеристиках $g(v)$ вершин, возможно не определенных для сравниваемых графов:

$$d_v\big((V_1, E_1), (V_2, E_2)\big) = \begin{cases} 0, v \notin V_1, v \notin V_2, \\ \lambda_1\sqrt{2}\sqrt{1 - \cos(\pi\lambda_2 \frac{g_1 - g_2}{\sup(g) - \inf(g)})}, v \in V_1, v \in V_2, \\ \lambda_1 \text{ иначе}, \end{cases}$$

где $\lambda_1, \lambda_2$ — параметры функции $d_v$.

## 1.4. Порождениие и выбор структуры модели глубокого обучения

В данном разделе рассматриваются работы, посвященные порождению и модификации структуры моделей. В отличие от работ, описанных в предыдущих разделах, в следующих работах рассматриваемым объектом является не отдельный параметр, а подмодель или группа параметров, входящая в эту подмодель.

**Графовое представление структуры модели.** Одним из возможных представлений структуры моделей глубокого обучения является графовое представление, в котором в качестве ребер графа выступают нелинейные функции, а в качестве вершин графа — представление выборки под действием соответствующих нелинейных функций. Данный подход к описанию модели является соответсвтует походам, описанным в [83], а также в библиотеках типа TensorFlow [84], Theano [85], Pytorch [86], в которых модель рассматривается как граф, ребрами которого выступают математические операции, а вершинами — результат их действия на выборку. В то же время, существуют и другие способы представления модели. В ряде работ, посвященных байесовской оптимизации [80, 79, 76], модель рассматривается как черный ящик, над которым производится ограниченный набор операций типа “произвести оптимизацию параметров” и “предсказать значение зависимой переменной по независимой пе-

ременной и параметрам модели". Подход, описанный в данных работах, также коррелирует с библиотеками машинного обучения, такими как Weka [87], RapidMiner [88] или sklearn [89], в которых модель машинного обучения рассматривается как черный ящик.

В [90] представлен обзор по графовому описанию моделей глубокого обучения, предлагается метод формального описания графовых сетей (англ. Graph Network), являющийся обобщением предложенных ранее графовых описаний моделей.

В работе [91] рассматриваются подходы к порождению моделей глубокого обучения. Предлагается формализация пространства поиска и формальное описание элементов пространства моделей. Приведем пример описания параметрического семейства моделей, соответствующего схеме из Рис. 1.1 при условии, что структурные параметры $\boldsymbol{\gamma}$ имеют только одну ненулевую компоненту:

```
(Concat
  OR(
    (Conv2D [c_0] [c_1] [1],
    (Concat(
      (Conv2D [c_0] [c_2] [1],
      (Conv2D [1] [c_1] [1])),
  (Affine [10]),
  (SoftMax)).
```

**Прогнозирование графовых структур.** В работе [92] предлагается метод прогнозирования графовой структуры на основе линейного программирования. Предлагается свести проблему поиска графовой структуры к комбинаторной проблеме. В работе [93] предлагается метод прогнозирования структур деревьев, основанный на дважды-рекуррентных нейросетях (англ. doubly-reccurent), т.е. на сетях, отдельно прогнозирующих глубину и ширину уровней деревьев.

**Стохастическое порождение структур.** Одним из возможных методов порождения структр моделей глубокого обучения выступает стохастическое порождение структур. Данный тип порождения предполагает, что структуры порождаются случайно в соответствие вариационным распределением, заданным на структурах $q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$. Затем выбирается одна, либо несколько наилучших структур с учетом валидационной функции $Q$ или внешних, возможно недифференцируемых, критериев качества. Итоговая модель получается путем оптимизации параметров модели при выбранной структуре $\mathbf{\Gamma}$. Заметим, что в ряде работ, одновременно порождается не только структура модели, но и итоговые параметры.

В работе [94] рассматривается порождение моделей, оптимизируемых без учителя. Модель представляется многослойным перцептроном вида:

$$\mathbf{f} = \mathbf{f}_{|\mathrm{V}|\text{-}1} \circ \cdots \circ \mathbf{f}_0(\mathbf{x}), \quad \mathbf{f}_i(\mathbf{x}) = \boldsymbol{\sigma}(\mathbf{x}\big(\mathbf{w}^i \odot \mathbf{H}^i\big)),$$

где $\mathbf{H}^i$ — бинарные матрицы, определяющие вклад каждого параметра из $\mathbf{w}^i$ в итоговую модель, знаком $\odot$ обозначается покомпонентное перемножение.

Порождение моделей производится с использованием композиции процессов индийских буфетов. Процесс индийского буфета заключается в итеративном построении матрицы $\mathbf{H}^i$ с ограниченным, но не заданным наперед количеством столбцов. Интепретируя количество столбцов матрицы как размер $i$-го слоя предлагается метод, позволяющий выбирать стохастически порождать модели с разилчной размерностью скрытых слоев.

В работе [95] предлагается метод выбора модели сверточной нейронной сети. Используется функция потерь, основанная на аппроксимации априорного распределения процесса индийского буфета для каждой базовой функции $\mathbf{g}_j$, являющейся $j$-м отображением объектов:

$$L = \sum_{\mathbf{x}\in\mathbf{X}} ||\mathbf{x} - \sum_{j=0}^{K-1} \mathbf{w}^j * \mathbf{g}_j(\mathbf{x})||_2^2 + \lambda^2 K,$$

где $K$ — параметр, отвечающий за количество фильтров, $\lambda$ — метапараметр алгоритма, знаком $*$ обозначается операция свертки.

В работе [96] предлагается ввести априорное распределение Бернулли на структурные параметры $\boldsymbol{\gamma}^{j,k} \in \Gamma$.

В [97] рассматривается задача выбора архитектуры с помощью большого количества параллельных запусков обучения моделей. Предлагаются критерии ранней остановки процедуры оптимизации обучения моделей.

**Последовательный выбор структуры модели.** В работе [5] приводятся теоретические оценки построения нейросетей с использованием жадных стратегий, при которых построение модели производится итеративно последовательным увеличением числа нейронов в сети. В работе [6] предлагается жадная стратегия выбора модели нейросети с использованием релевантных априорных распределений, т.е. параметрических распределений, оптимизация параметров которых позволяет удалить часть параметров из модели. Данный метод был к задаче построения модели метода релевантных векторов [7].

В работах[11, 12] рассматривается послойное построение модели с отдельным критерием оптимизации для каждого слоя. В работах [13, 14, 15] предлагается декомпозиция модели на порождающую и разделяющую, оптимизируемых последовательно.

В работах [98, 16] предлагается наращивание моделей, основанное на бустинге. Рассматривается задача построения нейросетевых моделей специального типа:

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}_{|V|-1} \circ \mathbf{f}_{|V|-2} \circ \dots \mathbf{f}_0(\mathbf{x}), \quad \mathbf{f}_{i+1}(\mathbf{x}) = \boldsymbol{\sigma}\left(\mathbf{f}_i(\mathbf{x})\right) + \mathbf{f}_i(\mathbf{x}),$$

приводится параметризация модели, позволяющая рассматривать декомпозировать модель на слабые классификаторы. В [16] рассматривается задача выбора полносвязной нейронной сети для задачи бинарной классификации, $R = 2$.

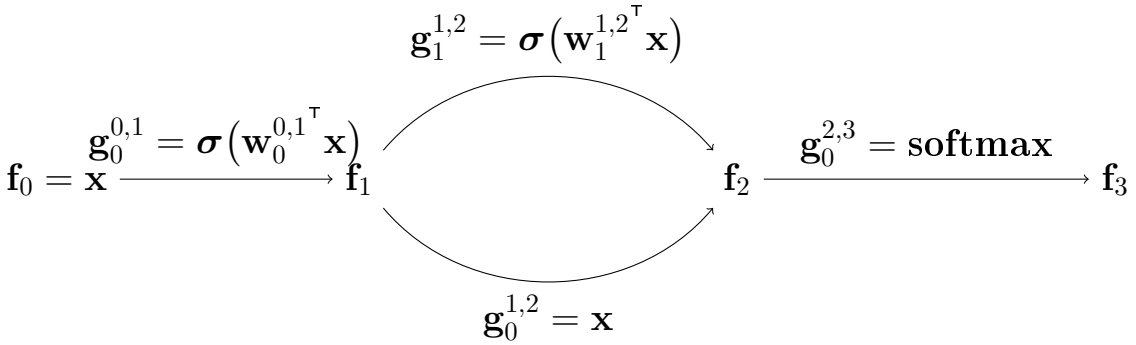



Рис. 1.4. Пример итерации алгоритма AdaNet [16]. Рассматривается две альтернативные модели: модель с углублением сети (соответствует занулению функции $\mathbf{f}_2$ с использованием базовой функции $\mathbf{g}_1^{1,2}$) и модель с расширением сети (соответствует базовой функции $\mathbf{g}_0^{1,2}$).

В качестве функции агрегации для подмодели $\mathbf{f}_3$ выступает конкатенация: $\mathbf{agg}_2 = \mathbf{concat}$.

На каждом шаге построения выбирается одно из двух расширений модели, каждое из которых рассматривается как слабый классификатор: сделать модель шире или сделать модель глубже. Пример работы AdaNet представлен на Рис. 1.4. Построение модели заканчивается при условии снижении радемахеровской сложности:

$$\mathfrak{R} = \frac{1}{m}\mathsf{E}_{b_1,\ldots,b_m}\sup_{\mathbf{w}}\sum_{i=1}^{m} b_i[y_i \neq \arg\max_c f[c](\mathbf{x},\mathbf{w})], \tag{1.14}$$

где $b_i$ — реализация случайной дискретной величины, равновероятно принимающей значений $-1$ и 1, $f[c]$ — $c$-я компонента модели $\mathbf{f}$.

В работе [99] рассматривается задача порождения сверточных нейронных сетей. Предлагается проводить последовательный выбор структуры модели по восходящему числу параметров: начиная от сетей с одной подмоделью и итеративно увеличивая количество подмоделей. В силу высокой вычислительной сложности данного подхода, вместо последовательного порождения моделей, предлагается провести оптимизацию рекуррентной нейронной сети, которая предсказывает качество модели по заданным подмоделям, и на основе данного предсказания выбрать наилучшую модель.

В работе [100] предлагается метод анализа структуры сети на основе линейных классификаторов, построенных на промежуточных слоях нейросети. Схожий метод был предложен в [101], где классификаторы на промежуточных уровнях используются для уменьшения вычислений при выполнении вывода и предсказаний. Промежуточные классификаторы работают как решающий список.

В работе [102] предлагается инкрементальный метод оптимизации нейросети. На первом этапе модель декомпозируется на несколько подмоделей, при

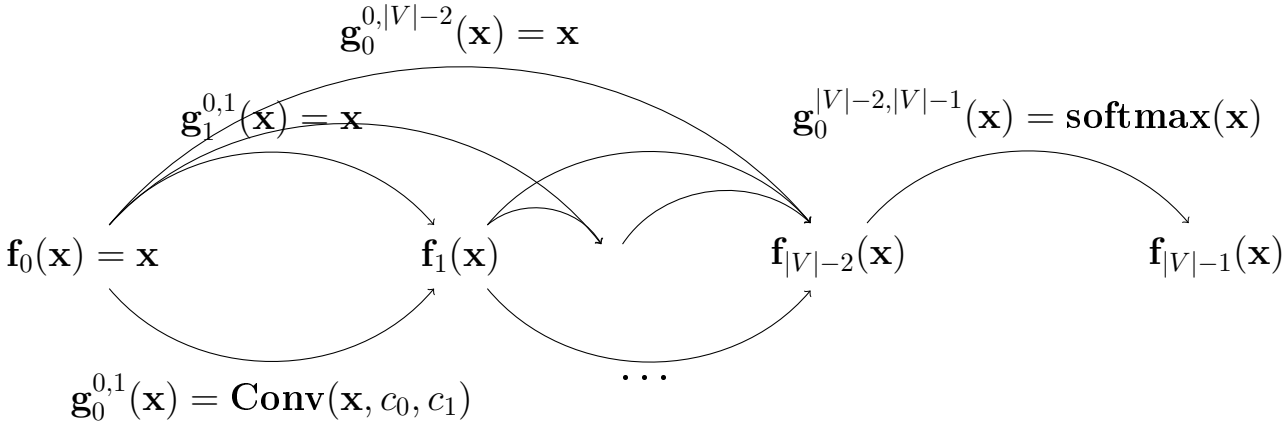



Рис. 1.5. Пример параметрического семейства моделей глубокого обучения, описываемый в [17]. Каждая подмодель $\mathbf{f}_j$ является линейной комбинацией базовых функций: свертки и результата работы предыдущих подмоделей (англ. skip-connection).

которой модель последовательностью слоев $\mathbf{f}_1, \ldots, \mathbf{f}_{|V|}$. Проводится последовательная оптимизация моделей вида:

1) $\mathbf{f} = \mathbf{f}_{|V|-1}(\mathbf{x})$;
2) $\mathbf{f} = \mathbf{f}_{|V-2|} \circ \mathbf{f}_{|V|}(\mathbf{x})$;
3) ...
4) $\mathbf{f} = \mathbf{f}_0 \circ \cdots \circ \mathbf{f}_{|V-1|}(\mathbf{x})$.

**Оптимизация структуры модели на основе обучения с подкреплением.** В [17] предлагается итеративная схема выбора архитектуры сверточной нейросети с использованием обучения с подкреплением. Распределение структур и параметров $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})$ задается рекуррентной нейронной сетью, которая определяет значение параметров модели и наличие ребер с ненулевыми операциями между вершинами графов модели. Параметры рекуррентной нейронной сети оптимизируются на основе значения функции $Q$, получаемого на каждой итерации алгоритма.

В работе [18] предлагается алгоритм построения регрессионной модели для оценки финального качества модели и ранней остановки оптимизации моделей. Он позволяет существенно ускорить поиск моделей, представленный в [17]. В [20] рассматривается задача переноса архитектуры нейросети, чья структуры была выбрана по выборке, меньшей мощности. Как и в [17] предлагается метод параметризации сверточной нейронной сети в виде графа. Предложенная параметризация позволяет задать более мощное параметрическое семейство моделей, чем в [17]. Модель представляется в виде последовательности суперпозиций подмоделей, называемых клетками (англ. normal cell и reduction cell). Каждая из этих клеток содержит следующее множество нелинейных операций $\mathbf{g}$, состоящее из тождественной операции $\mathbf{g}(\mathbf{x}) = \mathbf{x}$, а также множество сверток

с фиксированным количеством каналов и размером фильтров и функций субдискретизации или пулинга. Алгоритм выбора структуры модели рекуррентной сетью выглядит следующим образом на шаге $j$:

1) выбрать вершину $v'$ из вершин $v_{j-1}$, $v_{j-2}$ из данной клетки или вершину из предыдущих клеток;
2) выбрать вершину $v''$ из вершин $v_{j-1}$, $v_{j-2}$ из данной клетки или вершину из предыдущих клеток;
3) выбрать базовую функцию $\mathbf{g}'$ для применения к вершине $v'$;
4) выбрать базовую функцию $\mathbf{g}''$ для применения к вершине $v''$;
5) выбрать функцию агрегации результатов применения операций $\mathbf{g}', \mathbf{g}''$: сумму или конкатенацию.

В отличие от предыдущих работ, в работе [19] предлагается подход к инкрементальному обучению нейросети, основанном на модификации модели, полученной на предыдущем шаге. Рассматривается две операции над нейросетью: расширение и углубление сети.

В работах [103, 104, 105] рассматриваются методы деформации нейросетей. В работе [105] предлагается метод оптимального разделения нейросети на несколько независимых сетей для уменьшения количества связей и, как следствие, уменьшения сложности оптимизации модели. В работе [103] предлагается метод сохранения результатов оптимизации нейросети при построении новой более глубокой или широкой нейросети. В работе [104] рассматривается задача расширения сверточной нейросети, нейросеть рассматривается как граф.

В работе [21] используется представление модели из [20]. Вместо обучения с подкреплением испольщуются градиентная оптимизация структуры и параметров, выполненная в единой процедуре.

## 1.5. Метаоптимизация моделей глубокого обучения

Задача выбора структуры модели тесно связана с раздел машинного обучения под названием *метаобучение* или *метаоптимизация*. Под метаобучением понимаются алгоритмы машинного обучения [106], которые:

1) оценивают и сравнивают методы оптимизации моделей;
2) оценивают возможные декомпозиции процесса оптимизации моделей;
3) на основе полученных оценок предлагают оптимальные стратегии оптимизации моделей и отвергают неоптимальные.

В работе [107] предлагается подход к адаптивному изменению параметров сети. В качестве оператора оптимизации параметров рассматривается величина:

$$T(\boldsymbol{\theta}|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \boldsymbol{\theta} + \mathbf{f}_{\text{optim}}(\mathbf{f}_{\text{mod}}(\boldsymbol{\theta})),$$

где $\mathbf{f}_{\text{mod}}$ — функция, определяющая номер параметра из $\boldsymbol{\theta}$, подлежащего оптимизации, а $\mathbf{f}_{\text{optim}}$ — величина изменения параметра. В [107] также предлагается подмодель $\mathbf{f}_{\text{ana}}$, определяющая номер параметра, подлежащего дальнейшему

анализу. Подход, описанный в данной работе, предполагает оптимизацию и анализ не только самой модели $\mathbf{f}$, но и дополнительных моделей $\mathbf{f}_{\text{mod}}, \mathbf{f}_{\text{ana}}, \mathbf{f}_{\text{optim}}$.

В работе [108] рассматривается оптимизация метапараметров (шага градиентного спуска $\lambda_{\text{lr}}$ и начального распределения параметров $\boldsymbol{\theta}^0$). Рассматривается задача оптимизации параметров модели в случае, когда количество примеров невелико. Для этого проводится оптимизация параметров оператора оптимизации, который выглядит следующим образом:

$$T(\boldsymbol{\theta}|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \boldsymbol{\theta}^0 - \boldsymbol{\lambda}\nabla T(\boldsymbol{\theta}^0|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}),$$

где векторы $\boldsymbol{\theta}^0$ и $\boldsymbol{\lambda}$ являются метапараметрами оператора $T$. Задача оптимизации параметров оператора $T$ рассматривается как задача многзадачного обучения (англ. multitask learning), когда оператор оптимизируется с учетом нескольких различных выборок и различных функций $L$, определенных отдельно для каждой выборки.

В работе [109] рассматирвается задача восстановления параметров модели по параметрам другой модели, чьи параметры были получены оптимизацией функции потерь на выборке меньшей мощности. Задачу можно рассматривать как задачу нахождения параметров некоторого оператора оптимизации $T$, действующего на параметры $\boldsymbol{\theta}^0$, где $\boldsymbol{\theta}^0$ — параметры модели, оптимизированной на небольшой выборке. Предлагается функция оптимизации:

$$T(\boldsymbol{\theta}|L, \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \arg\min ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0||_2^2 + \lambda L(\boldsymbol{\theta}|\hat{\mathbf{y}}, \hat{\mathbf{X}}, \mathbf{h}, \boldsymbol{\lambda}),$$

где $\boldsymbol{\theta}$ — параметры модели, обученной по полной выборке $\mathfrak{D}$, $\hat{\mathfrak{D}}$ — выборка меньшей мощности, $\lambda$ — настраиваемый метапараметр.

В работе [110] рассматривается оптимизация метапараметров оператора оптимизации с помощью модели долгой краткосрочной памяти LSTM, которая выступает альтернативе аналитических алгоритмов, таких как Adam [111] или AdaGrad [112]. LSTM имеет небольшое число параметров, т.к. для каждого метапараметра используется свой экземпляр модели LSTM с одинаковыми параметрами для каждого экземпляра. Оптимизируемый функционал является суммой значений функции потерь $L$ на нескольких шагах оптимизации:

$$Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{t=1}^{\eta} L(\boldsymbol{\theta}^t|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}),$$

где $\eta$ — число шагов оптимизации, $\boldsymbol{\theta}^t$ — оптимизируемые параметры модели на шаге оптимизации $t$.

## 1.6. Выбор структур моделей специального вида

В данном разделе представлены работы по поиску оптимальных моделей со структурами специального вида.

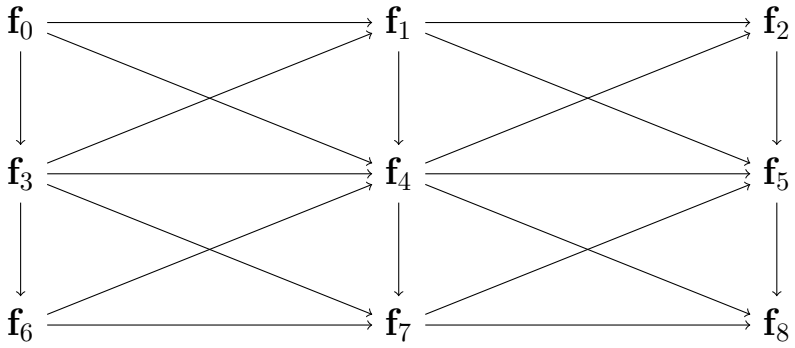



Рис. 1.6. Пример суперсети. Каждый путь из подмодели $\mathbf{f}_0$ в конечную модель $\mathbf{f}_8$ задает модель глубокого обучения.

В работе [113] рассматривается оптимизация моделей нейросетей с бинарной функцией активацией. Задача оптимизации сводится к задаче mixed integer программирования, которая решается методами выпуклого анализа. В работе [114] предлагается метод построения сети глубокого обучения, структура которой выбирается с использованием обучения без учителя. Критерий оптимальности модели использует оценки энергитических функций и ограниченной машины Больцмана.

В работах [115, 116] рассматривается выбор архитектуры сети с использованием *суперсетей*: связанных между собой подмоделей, образующих граф, каждый путь из нулевой вершины в последнюю которого определяет модель глубокого обучения. Пример графа, описывающего суперсеть представлен на Рис. 1.6. В работе [116] рассматриваются стохастические суперсети, позволяющие выбрать структуру нейросети за органиченное время оптимизации. Схожий подход был предложен в работе [115], где предлагается использовать эволюционные алгоритмы для запоминания оптимальных подмоделей и переноса этих моделей в другие задачи.

**Порождающие модели.** Порождающими моделями называются модели, приближающие совместное распределение объектов и соответствующих им меток $p(\mathbf{X}, \mathbf{y})$. Частным случаем порождающих моделей являются модели, приближающие только распределение векторов объектов $\mathbf{X}$. Подобный случай будем считать частным случаем классификации при пустом множестве меток классов ($R = 0$).

В качестве порождающих моделей в сетях глубокого обучения выступают ограниченные машины Больцмана [3] и автокодировщики [22]. В работе [23] рассматриваются алгоритмы регуляризации автокодировщиков, позволяющих формально рассматривать данные модели как порождающие модели с использованием байесового вывода. В работе [24] рассматриваются регуляризованные автокодировщики и свойства оценок их правдоподобия. В работе [25] предлагается обобщение автокодировщика с использованием вариационного байесовского вывода [2]. В работе [26] рассматриваются модификации вариационного

автокодировщика и ступенчатых сетей [27] для случая построения многослойных порождающих моделей.

В ряде работ [117, 118, 119, 120, 121] рассматривается подход к построению порождающих моделей глубокого обучения, при котором каждая подмодель $\mathbf{f}_i$ приближает распределение некоторой случайной величины $\mathbf{z}_i$, которая влияет на маргинальное распределение $p(\mathbf{X}, \mathbf{y}) = \int_{\mathbf{z}_0,\ldots,\mathbf{z}_{|V|-1}} p(\mathbf{X}, \mathbf{y}|\mathbf{z}_0, \ldots, \mathbf{z}_{|V|-1}) p(\mathbf{z}_1, \ldots, \mathbf{z}_{|V|}) d\mathbf{z}_0 \ldots d\mathbf{z}_{|V|-1}$. Подобный подход позволяет использовать вероятностную интерпретацию для каждой отдельной подмодели.

В работе [117] рассматривается обобщение вариационного автокодировщика на случай более общих графических моделей. Предлагается проводить оптимизацию сложных графических моделей в единой процедуре. Для вывода предлагается использовать нейронные сети. Другая модификация вариационного автокодировщика представлена в работе [118], авторы рассматривают использование процесса сломанной трости в вариационном автокодировщике, тем самым получая модель со стохастической размерностью скрытой переменной. В [119] рассматривается смесь автокодировщиков, где смесь моделируется процессом Дирихле.

В работе [120] предлагается подход к оптимизации неизвестного распределения с помощью вариационного вывода. Предлагается решать задачу оптимизации итеративно, добавляя в модель новые компоненты вариационного распределения, проводится аналогия с бустингом.

В работе [121] рассматривается задача построения порождающих моделей с дискретными значениями скрытой переменной $\mathbf{z}$, предлагается критерий для послойного обучения порождающих моделей:

$$Q = \sum_{\mathbf{x} \in \mathbf{X}} \log \sum_i p(\mathbf{x}|\mathbf{z}_i) q(\mathbf{z}) \to \max,$$

где $q$ — аппроксимирующее распределение для случайной величины $\mathbf{z}$, $i$ пробегает все значения переменной $\mathbf{z}$.

В работе [53] рассматривается метод ARD для снижения размерности скрытого пространства вариационных порождающих моделей. Скрытая переменная параметризуется как произведение некоторой случайной величины $\mathbf{z}$ на вектор $\mathbf{h}$, отвечающий за релевантность каждой компоненты скрытой переменной. Схема порождения выборки $\mathbf{X}$ представлена на Рис. 1.7.

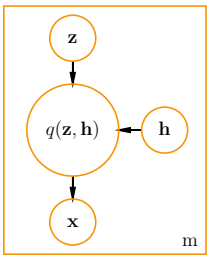



Рис. 1.7. Схема порождения вектора объектов $\mathbf{X}$, представленная в [53].

В данной работе предлагается метод последовательного порождения моделей глубокого обучения, основывающийся на применении вариационного вывода. Вариационный ввывод позволяет получить оценки правдоподобия модели с небольшими вычислительными затратами, а также проследить потенциальное начало переобучения модели без использования контрольной выборки. Для регуляризации структуры модели предлагается ввести априорное распределение на структуре, позволяющее проводить оптимизацию модели и ее структуры в различных режимах. В качестве метода оптимизации гиперпараметров выступают градиентные методы, что позволяет эффективно производить оптимизацию большого числа гиперпараметров, сопоставимого с числом параметров модели.

# Глава 2
# Выбор субоптимальной структуры модели

В данной главе рассматривается задача выбора структуры модели глубокого обучения. Предлагается ввести вероятностные предположения о распределении параметров и распределении структуры модели. Проводится градиентная оптимизация параметров и гиперпараметров модели на основе байесовского вариационного вывода. 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. Решается двухуровневая задача оптимизации: на первом уровне проводится оптимизация нижней оценки обоснованности модели по вариационным параметрам модели. На втором уровне проводится оптимизация гиперпараметров модели.

## 2.1. Вероятностная модель

Определим априорные распределения параметров и структуры модели следующим образом. Пусть для каждого ребра $(j, k) \in E$ и каждой базовой функ-

ции $\mathbf{g}_l^{j,k}$ 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, $l \in \{1, \ldots, K^{j,k}\}$, где $K^{j,k}$ — количество базовых функций для ребра $K^{j,k}$. Априорное распределение $p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h})$ 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 $\mathbf{\Gamma}$ предлагается использовать произведение распределений Gumbel-Softmax ($\mathcal{GS}$) [122]:

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

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:

$$p(\boldsymbol{\gamma}^{j,k}|\mathbf{s}^{j,k}, \lambda_{\text{temp}}) = (K^{j,k}-1)!(\lambda_{\text{temp}})^{K^{h,j}-1} \prod_{l=1}^{K^{j,k}} s_l^{j,k} (\boldsymbol{\gamma}_l^{j,k})^{-\lambda_{\text{temp}}-1} \times \\ \times \left( \sum_{l=1}^{K^{j,k}} s_l^{j,k} (\boldsymbol{\gamma}_l^{j,k})^{-\lambda_{\text{temp}}} \right)^{-K^{j,k}}, \quad (2.1)$$

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.

Перечислим свойства, которыми обладает распределение Gumbel-Softmax:

1. Компонента $l$ случайной величины $\boldsymbol{\gamma}^{j,k}$ представима следующим образом:

$$\boldsymbol{\gamma}_l^{j,k} = \frac{\exp(\log s_l^{j,k} + G_l^{j,k})/\lambda_{\text{temp}}}{\sum_{l'=1}^{K^{j,k}} \exp(\log s_{l'}^{j,k} + G_{l'}^{j,k})/\lambda_{\text{temp}}}, \quad (2.2)$$

   где $\mathbf{G}^{j,k} \sim -\log\big(-\log \mathcal{U}(0,1)^{K^{j,k}}\big)$.

2. Свойство округления: $p(\boldsymbol{\gamma}_{l_1} > \boldsymbol{\gamma}_{l_2}, l_1 \neq l_2 | \mathbf{s}^{j,k}, \lambda_{\text{temp}}) = \frac{s_{l_1}^{j,k}}{\sum_{l'} s_{l'}^{j,k}}$.

3. При устремлении температуры к нулю плотность случайной величины концентрируется на вершинах симплекса:

$$p(\lim_{\lambda_{\text{temp}}\to 0} \gamma_l^{j,k} = 1 | \mathbf{s}^{j,k}, \lambda_{\text{temp}}) = \frac{s_l^{j,k}}{\sum_{l'} s_{l'}^{j,k}}.$$

4. При устремлении температуры к бесконечности плотность распределения концентрируется в центре симплекса:

$$\lim_{\lambda_{\text{temp}}\to\infty} p(\boldsymbol{\gamma}^{j,k}|\mathbf{s}^{j,k},\lambda_{\text{temp}}) = \begin{cases}\infty, \boldsymbol{\gamma}^{j,k} = \frac{1}{K^{j,k}}, l \in \{1,\dots,K^{j,k}\},\\ 0, \text{ иначе.}\end{cases} \tag{2.3}$$

Доказательства первых трех утверждений приведены в [122]. Докажем утверждение 4.

*Доказательство.* Формула плотности с точностью до множителя записывается следующим образом :

$$p(\boldsymbol{\gamma}^{j,k}|\mathbf{s}^{j,k},\lambda_{\text{temp}}) \propto \frac{(\lambda_{\text{temp}})^{K^{j,k}-1}}{\left(\sum_{l=1}^{K^{j,k}} s_l^{j,k}(\boldsymbol{\gamma}_l^{j,k})^{-\frac{K^{j,k}-1}{K^{j,k}}\lambda_{\text{temp}}} \prod_{l'=1}^{K^{j,k}}[l\neq l'](\boldsymbol{\gamma}_l^{j,k})^{\frac{1}{K^{j,k}}\lambda_{\text{temp}}}\right)^{K^{j,k}}}. \tag{2.4}$$

Заметим, что числитель $(\lambda_{\text{temp}})^{K^{j,k}-1}$ имеет меньшую скорость сходимости, чем знаменатель, поэтому для вычисления предела достаточно проанализировать только знаменатель. Знаменатель под степенью $(K^{j,k})$ представляется суммой слагаемых следующего вида:

$$\left(\frac{\prod_{l'\neq l}\gamma_{l'}^{\frac{1}{K^{j,k}}}}{\gamma_l^{\frac{K^{j,k}-1}{K^{j,k}}}}\right)^{\lambda_{\text{temp}}}. \tag{2.5}$$

Рассмотрим два случая: когда вектор $\boldsymbol{\gamma}^{j,k}$ лежит не в центре симплекса, и когда $\boldsymbol{\gamma}^{j,k}$ лежит в центре симплекса. Пусть хотя бы для одной компоненты $l$ выполнено: $\boldsymbol{\gamma}_l^{j,k} \neq \frac{1}{K^{j,k}}$. Пусть $l'$ соответствует индексу максимальной компоненты вектора $\boldsymbol{\gamma}^{j,k}$:

$$l' = \underset{l\in\{1,\dots,K^{j,k}\}}{\arg\max}\ \boldsymbol{\gamma}_l^{j,k}.$$

Для $l = l'$ предел выражения (2.5) при $\lambda_{\text{temp}} \to \infty$ стремится к бесконечности. Для $l \neq l'$ предел выражения (2.5) при $\lambda_{\text{temp}} \to \infty$ стремится к нулю. Возводя сумму пределов в степень $(-K^{j,k})$ получаем предел плотности, равный нулю.

Рассмотрим второй случай. Пусть $\gamma_l^{j,k} = \frac{1}{K^{j,k}}$ для всех компонент вектора $\boldsymbol{\gamma}^{j,k}$. Тогда выражение (2.1) с точностью до множителя упрощается до $(\lambda_{\text{temp}})^{K^{j,k}-1}$. Предел данного выражения стремится к бесконечности. Таким образом, предел плотности Gumbel-Softmax равен выражению (2.3), что и требовалось доказать.

□

Первое свойство Gumbel-Softmax распределения позволяет использовать репараметризацию при вычислении градиента в вариационном выводе (англ. reparametrization trick).

**Определение 20.** Случайную величину $\psi$ с распределением $q$ с параметрами $\boldsymbol{\theta}_\psi$ назовем репараметризованной через случайную величину $\varepsilon$, чье распределение не зависит от параметров $\boldsymbol{\theta}_\psi$, если:

$$\psi = g(\varepsilon, \boldsymbol{\theta}_\psi)$$

где $g$ — некоторая непрерывная функция.

Идею репараметризации поясним на следующем примере.

**Пример 2.** Пусть структура $\boldsymbol{\Gamma}$ зафиксирована для модели $\mathbf{f}$. Рассмотрим математическое ожидание логарифма правдоподобия выборки модели по некоторому непрерывному распределению $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$:

$$\mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} \log\ p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) = \int_{\mathbf{w}} \log\ p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) d\mathbf{w}.$$

Продифференцируем данное выражение по параметрам $\boldsymbol{\theta}_{\mathbf{w}}$ вариационного распределения $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$, полагая что оно удовлетворяет необходимым условиям для переноса оператора дифференцирования под знак интеграла:

$$\nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) = \int_{\mathbf{w}} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) \nabla_{\boldsymbol{\theta}_{\mathbf{w}}} q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) d\mathbf{w}.$$

Это выражение в общем виде не имеет аналитического решения. Пусть распределение $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ для параметров $\mathbf{w}$ подлежит репараметризации через случайную величину $\boldsymbol{\varepsilon}$:

$$\mathbf{w} = \mathbf{g}(\boldsymbol{\varepsilon}, \boldsymbol{\theta}_{\mathbf{w}}).$$

Тогда справедливо следующее выражение:

$$\nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) = \nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathsf{E}_{\boldsymbol{\varepsilon}} \log p(\mathbf{y}|\mathbf{X}, \mathbf{g}(\boldsymbol{\varepsilon}), \boldsymbol{\Gamma}) =$$

$$= \int_{\boldsymbol{\varepsilon}} \nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \log p(\mathbf{y}|\mathbf{X}, \mathbf{g}(\boldsymbol{\varepsilon}), \boldsymbol{\Gamma}) p(\boldsymbol{\varepsilon}) d\boldsymbol{\varepsilon} = \mathsf{E}_{\boldsymbol{\varepsilon}} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{y}|\mathbf{X}, \mathbf{g}(\boldsymbol{\varepsilon}), \boldsymbol{\Gamma}).$$

Таким образом, распределение, позволяющее произвести репараметризацию, является более удобным для вычисления интегральных оценок вида $\nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma})$, а также позволяет повысить точность приближенного вычисления значений таких функций [123]. Подробный анализ репараметризации для генеративных моделей глубокого обучения представлен в [124].

Пример распределения Gumbel-Softmax при различных параметрах представлен на Рис. 2.1. В качестве альтернативы для априорного распределения структуры выступает распределение Дирихле. В качестве предельного случая, когда все структуры $\boldsymbol{\Gamma} \in \mathbb{\Gamma}$ равнозначны, выступает равномерное распределение. Выбор в качестве распределения стрктуры произведения распределений Gumbel-Softmax обоснован выбором этого распределения в качестве вариационного.

(а) (б) (в) (г)

Рис. 2.1. Пример распределения Gumbel-Softmax при различных значениях параметров: а) $\lambda_{\text{temp}} \to 0$, б) $\lambda_{\text{temp}} = 1, \mathbf{s} = [1, 1, 1]$, в) $\lambda_{\text{temp}} = 5, \mathbf{s} = [1, 1, 1]$, г) $\lambda_{\text{temp}} = 5, \mathbf{s} = [10, 0.1, 0.1]$.

Заметим, что предлагаемое априорное распределение неоднозначно: одно и то же распределение можно получить с различными значениями гиперпараметра $\mathbf{A}_l^{j,k}$ и структурного параметра $\gamma_l^{j,k}$. 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. Использование обратного гамма-распределения в качестве распределения гиперпараметров можно найти в [2, 28]. В данной работе обратное распределение выступает как регуляризатор гиперпараметров. Варьированием метапараметров $\lambda_1, \lambda_2$ получается более сильная или более слабая регуляризация [7]. Пример распределений $\text{inv-gamma}(\lambda_1, \lambda_2)$ для разных значений метапараметров $\lambda_1, \lambda_2$ изображен на Рис. 2.2. Оптимизации без регуляризации соответствует случай предельного распределения $\lim_{\lambda_1, \lambda_2 \to 0} \text{inv-gamma}(\lambda_1, \lambda_2)$.

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.

График вероятностной модели в формате плоских нотаций представлен на Рис. 2.3.

## 2.2. Вариационная оценка обоснованности вероятностной модели

Задача выбора структуры $\boldsymbol{\Gamma}$ и параметров $\mathbf{w}$ заключается в получении оценок на апостериорное распределение $p(\mathbf{w}, \boldsymbol{\Gamma} | \mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) =$

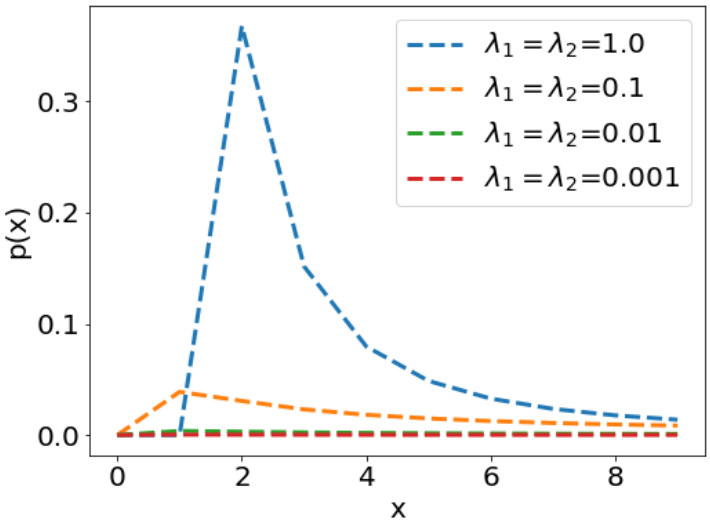



Рис. 2.2. Графики обратных гамма распределений для различных значений метапараметров.

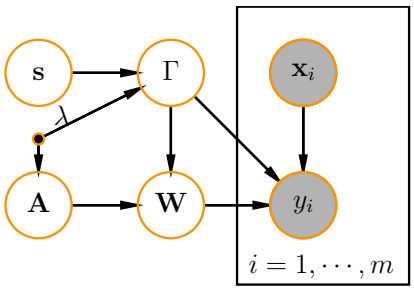



Рис. 2.3. График предлагаемой вероятностной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Наблюдаемые переменные обозначены серыми кругами.

$p(\mathbf{\Gamma}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})p(\mathbf{w}|\mathbf{y},\mathbf{X},\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})$. Оно зависит от гиперпараметров $\mathbf{h}$. В качестве критерия выбора гиперпараметров предлагается использовать апостериорную вероятность гиперпараметров:

$$p(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\lambda}) \propto p(\mathbf{y}|\mathbf{X},\mathbf{h},\boldsymbol{\lambda})p(\mathbf{h}|\boldsymbol{\lambda}) \to \max_{\mathbf{h}\in\mathbb{H}}. \tag{2.6}$$

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 (2.6).

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 (2.6) предлагается использовать нижнюю вариационную оценку обоснованности.

**Теорема 1.** 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:

$$\log p(\mathbf{y}|\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) \geq \tag{2.7}$$

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

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

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

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

*Доказательство.* Перепишем обоснованность:

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

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

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

Используя неравенство Йенсена получим

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

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

Декомпозируем распределение $q$ по свойству условной дивергенции:

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

$$= D_{\mathrm{KL}}\big(q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})||p(\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) + \mathsf{E}_{\mathbf{\Gamma}\sim q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})} \mathsf{E}_{\mathbf{w}\sim q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} \log \left( \frac{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})} \right). \quad (2.8)$$

□

В качестве вариационного распределения $q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})$ предлагается использовать нормальное распределение, не зависящее от структуры модели $\mathbf{\Gamma}$:

$$q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}),$$

где $\mathbf{A}_q^{-1}$ — диагональная матрица с диагональю $\boldsymbol{\alpha}_q$.

В качестве вариационного распределения $q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ предлагается использовать произведение распределений Gumbel-Softmax. Конкатенацию параметров концентрации распределений обозначим $\mathbf{s}_q$. Его температуру, общую для всех структурных параметров $\boldsymbol{\gamma} \in \mathbf{\Gamma}$, обозначим $\theta_{\text{temp}}$. Вариационными параметрами распределения $q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta})$ являются параметры распределений $q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}), q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$:

$$\boldsymbol{\theta} = [\boldsymbol{\mu}_q, \boldsymbol{\alpha}_q, \mathbf{s}_q, \theta_{\text{temp}}].$$

График вероятностной вариационной модели в формате плоских нотаций представлен на Рис. 2.4. Для анализа сложности полученной модели введем понятие *параметрической сложности.*

**Определение 21.** 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:

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

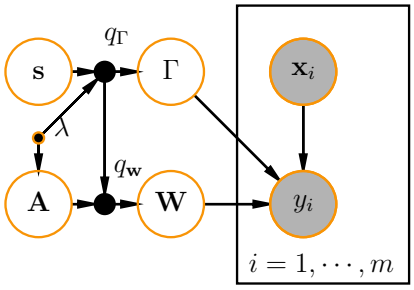



Рис. 2.4. График предлагаемой вероятностной вариационной модели в формате плоских нотаций. Переменные обозначены белыми и серыми кругами, константы обозначены обведенными черными кругами. Вариационное распределение обозначено черным кругом. Наблюдаемые переменные обозначены серыми кругами.

Параметрическая сложность модели соответствует минимальной по $\mathbf{h} \in U_\mathbf{h}$ минимальной ожидаемой длине описания параметров модели при условии заданного параметрического априорного распределения [125].

Одним из критериев удаления неинформативных параметров в вероятностных моделях является отношение вариационной плотности параметров в моде распределения к вариационной плотности параметра в нуле (1.9):

$$\frac{q_\mathbf{w}(w = \mu_q | \mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})}{q_\mathbf{w}(w = 0 | \mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})} = \exp\left(-\frac{\alpha_q^2}{2\mu_q^2}\right),$$

где параметру модели $w$ соответствуют вариационные параметры $\mu_q, \alpha_q$: $q_\mathbf{w}(w|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}) \sim \mathcal{N}(\mu_q, \alpha_q)$.

Обобщим понятие относительной вариационной плотности на случай произвольных непрерывных распределений.

**Определение 22.** Относительной вариационной плотностью параметра $w \in \mathbf{w}$ при условии структуры $\mathbf{\Gamma}$ и гиперпараметров $\mathbf{h}$ назовем отношение вариационной плотности в моде вариационного распределения параметра к вариационной плотности в моде априорного распределения параметра:

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

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

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

Сформулируем и докажем теорему о связи относительной плотности и параметрической сложности модели. Предварительно докажем две вспомогательные леммы.

**Лемма 1.** Пусть

1. Заданы компактные множества $U_{\mathbf{h}} \subset \mathbb{H}, U_{\boldsymbol{\theta}_\mathbf{w}} \subset \Theta_\mathbf{w}, U_{\boldsymbol{\theta}_\mathbf{\Gamma}} \subset \Theta_\mathbf{\Gamma}$.
2. Вариационное распределение $q_\mathbf{w}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})$ является абсолютно непрерывным и унимодальным на $U_{\boldsymbol{\theta}}$. Его мода и матожидание совпадают:
$$\text{mode } q_\mathbf{w}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}) = \mathsf{E}_{q_\mathbf{w}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})}\mathbf{w}.$$
3. Априорное распределение $p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ является абсолютно непрерывным и унимодальным на $U_\mathbf{h}$. Его мода и матожидание совпадают и не зависит от гиперпараметров $\mathbf{h}$ на $U_\mathbf{h}$ и структуры $\mathbf{\Gamma}$ на $U_{\boldsymbol{\theta}_\mathbf{\Gamma}}$:
$$\mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})}\ \mathbf{w} = \text{mode } p(\mathbf{w}|\mathbf{\Gamma}_1, \mathbf{h}_1, \boldsymbol{\lambda}) = \text{mode } p(\mathbf{w}|\mathbf{\Gamma}_1, \mathbf{h}_2, \boldsymbol{\lambda}) = \mathbf{m}$$
для любых $\mathbf{h}_1, \mathbf{h}_2 \in U_\mathbf{h}, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2 \in U_\mathbf{\Gamma}$.
4. Параметры модели $\mathbf{w}$ имеют конечные вторые моменты по маргинальным распределениям:
$$\int_\mathbf{\Gamma} q_\mathbf{\Gamma}(\mathbf{\Gamma}|\boldsymbol{\theta}_\mathbf{\Gamma}) q_\mathbf{w}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}) d\mathbf{\Gamma}, \quad \int_\mathbf{\Gamma} q_\mathbf{\Gamma}(\mathbf{\Gamma}|\boldsymbol{\theta}_\mathbf{\Gamma}) p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) d\mathbf{\Gamma}.$$
5. Вариационное распределение $q_\mathbf{w}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})$ является липшецевым по $\mathbf{w}$.
6. Значение $q_\mathbf{w}(\mathbf{m}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})$ не равно нулю при $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$.

Тогда
$$\left|\mathsf{E}_{q_\mathbf{\Gamma}(\mathbf{\Gamma}|\boldsymbol{\theta}_\mathbf{\Gamma})} \rho(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}, \mathbf{h}, \boldsymbol{\lambda}) - 1\right| \leq$$
$$\leq \frac{C_l}{\min_{\mathbf{\Gamma} \in \mathbb{\Gamma}, \boldsymbol{\theta}_\mathbf{w} \in U_{\boldsymbol{\theta}}} q_\mathbf{w}(\mathbf{m}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})} \iint_{\mathbf{\Gamma}, \mathbf{w}} |\mathbf{w}| \cdot |q_\mathbf{w}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}) - p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})| q_\mathbf{\Gamma}(\mathbf{\Gamma}|\boldsymbol{\theta}_\mathbf{\Gamma}) d\mathbf{w} d\mathbf{\Gamma},$$
где $C_l$ — максимальная константа Липшица для $q_\mathbf{w}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})$ на $U_{\boldsymbol{\theta}}$.

*Доказательство.* Для произвольного $\boldsymbol{\theta} = [\boldsymbol{\theta}_\mathbf{w}, \boldsymbol{\theta}_\mathbf{\Gamma}]$ рассмотрим выражение:
$$\left|\mathsf{E}_{q_\mathbf{\Gamma}(\mathbf{\Gamma}|\boldsymbol{\theta}_\mathbf{\Gamma})} \rho(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}, \mathbf{h}, \boldsymbol{\lambda}) - 1\right| =$$
$$\left| \int_\mathbf{\Gamma} \left( \frac{q_\mathbf{w}(\text{mode } q_\mathbf{w}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})}{q_\mathbf{w}(\text{mode } p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})} \right) q_\mathbf{\Gamma}(\mathbf{\Gamma}|\boldsymbol{\theta}_\mathbf{\Gamma}) d\mathbf{\Gamma} - 1 \right| =$$
*представляя единицу как дробь с равными знаменателем и числителем*
$$= \left| \int_\mathbf{\Gamma} \left( \frac{q_\mathbf{w}(\text{mode } q_\mathbf{w}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}) - q_\mathbf{w}(\text{mode } p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})}{q_\mathbf{w}(\text{mode } p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})} \right) q_\mathbf{\Gamma}(\mathbf{\Gamma}|\boldsymbol{\theta}_\mathbf{\Gamma}) d\mathbf{\Gamma} \right| =$$
*заменяя моду на матожидание (по условию теоремы)*
$$= \left| \int_\mathbf{\Gamma} \left( \frac{q_\mathbf{w}(\mathsf{E}_{q_\mathbf{w}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})}\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w}) - q_\mathbf{w}(\mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})}\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})}{q_\mathbf{w}(\mathbf{m}|\mathbf{\Gamma}, \boldsymbol{\theta}_\mathbf{w})} \right) q_\mathbf{\Gamma}(\mathbf{\Gamma}|\boldsymbol{\theta}_\mathbf{\Gamma}) d\mathbf{\Gamma} \right| \leq$$

*занося модуль под знак интеграла*

$$\leq \int_{\mathbf{\Gamma}} \left| \frac{q_{\mathbf{w}}(\mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}\mathbf{w})|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}) - q_{\mathbf{w}}(\mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})}\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\mathbf{m}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{\Gamma} \right| \leq$$

*используя липшецевость функции* $q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})$

$$\frac{C_l}{\min_{\mathbf{\Gamma}\in\mathbb{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}\in U_{\boldsymbol{\theta}}} q_{\mathbf{w}}(\mathbf{m}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} \int_{\mathbf{\Gamma}} |\mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}\mathbf{w} - \mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})}\mathbf{w}| q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{\Gamma} \leq$$

*расписывая матожидание через интеграл*

$$\leq \frac{C_l}{\min_{\boldsymbol{\theta}_{\mathbf{w}}\in U_{\boldsymbol{\theta}}} q_{\mathbf{w}}(\mathbf{m}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w}| \cdot |q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})| q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma},$$

что и требовалось доказать. □

**Лемма 2.** Пусть

1. Вариационное распределение $q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})$ и априорное распределение $p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})$ являются абсолютно непрерывными.
2. Решение задачи

$$\mathbf{h}^* = \arg\min_{\mathbf{h}\in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w},\mathbf{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\mathbf{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) \tag{2.9}$$

   единственно для любого $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$.
3. 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 нулю:

$$\iint_{\mathbf{w},\mathbf{\Gamma}} |p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h}[i],\boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}[i])| q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}[i])d\mathbf{\Gamma}d\mathbf{w},$$

   где $\boldsymbol{\theta}[i] = [\boldsymbol{\theta}_{\mathbf{w}}[i],\boldsymbol{\theta}_{\mathbf{\Gamma}}[i]]$ , $\mathbf{h}[i]$ — решение задачи (2.9) для $\boldsymbol{\theta}[i]$.

*Доказательство.* Воспользуемся неравенством Пинскера [**?**]:

$$||F_q(\boldsymbol{\theta}_{\mathbf{w}}[i]) - F_p(\mathbf{h}[i])||_{\mathrm{TV}} \leq \sqrt{\frac{1}{2}\widehat{\mathrm{KL}}\left(p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h}[i],\boldsymbol{\lambda})||q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}[i])\right)},$$

где $||\cdot||_{\mathrm{TV}}$ — расстояние по вариации, $F_q, F_p$ — функции распределения $q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}), p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})$, $\widehat{\mathrm{KL}}\left(p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})||q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})\right)$ — дивергенция при фиксированной структуре $\mathbf{\Gamma}$:

$$\int_{\mathbf{w}} q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}) \log\left(\frac{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})}\right) d\mathbf{w}.$$

По условию дивергенция (2.8) стремится к нулю при $i \to \infty$. Она декомпозируется на два неотрицательных слагаемых, поэтому оба они стремятся к нулю. Рассмотрим второе слагаемое:

$$0 = \lim_{i\to\infty} \mathsf{E}_{\boldsymbol{\Gamma}\sim q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i])} \mathsf{E}_{\mathbf{w}\sim q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}[i])} \log\left(\frac{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}[i])}{p(\mathbf{w}|\boldsymbol{\Gamma},\mathbf{h}[i],\boldsymbol{\lambda})}\right) =$$

*расписывая матожидание как интеграл*

$$= \lim_{i\to\infty} \left| \int_{\boldsymbol{\Gamma}} \int_{\mathbf{w}} \log\left(\frac{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}[i])}{p(\mathbf{w}|\boldsymbol{\Gamma},\mathbf{h}[i],\boldsymbol{\lambda})}\right) q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i]) q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}[i]) d\mathbf{w} d\boldsymbol{\Gamma} \right| \geq$$

*по неравенству Пинскера*

$$\geq \lim_{i\to\infty} \int_{\boldsymbol{\Gamma}} ||F_q(\boldsymbol{\theta}_{\mathbf{w}}[i]) - F_p(\mathbf{h}[i])||^2_{\mathrm{TV}} q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i]) d\boldsymbol{\Gamma} \geq 0.$$

Отсюда

$$\lim_{i\to\infty} \int_{\boldsymbol{\Gamma}} ||F_q(\boldsymbol{\theta}_{\mathbf{w}}[i]) - F_p(\mathbf{h}[i])||^2_{\mathrm{TV}} q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i]) d\boldsymbol{\Gamma} = 0.$$

По неравенству Йенсена

$$0 \leq \left( \int_{\boldsymbol{\Gamma}} ||F_q(\boldsymbol{\theta}_{\mathbf{w}}[i]) - F_p(\mathbf{h}[i])||_{\mathrm{TV}} q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i]) d\boldsymbol{\Gamma} \right)^2 \leq$$

$$\leq \int_{\boldsymbol{\Gamma}} ||F_q(\boldsymbol{\theta}_{\mathbf{w}}[i]) - F_p(\mathbf{h}[i])||^2_{\mathrm{TV}} q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i]) d\boldsymbol{\Gamma}.$$

Тогда по свойству степени предела

$$\lim_{i\to\infty} \int_{\boldsymbol{\Gamma}} ||F_q(\boldsymbol{\theta}_{\mathbf{w}}[i]) - F_p(\mathbf{h}[i])||_{\mathrm{TV}} q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i]) d\boldsymbol{\Gamma} = 0.$$

По лемме Шеффе [**?**] TODO данное выражение можно переписать как:

$$\lim_{i\to\infty} \frac{1}{2} \iint_{\mathbf{w},\boldsymbol{\Gamma}} |p(\mathbf{w}|\boldsymbol{\Gamma},\mathbf{h}[i],\boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}[i])| q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i]) d\boldsymbol{\Gamma} d\mathbf{w} = 0, \qquad (2.10)$$

что и требовалось доказать. □

**Теорема 2.** Пусть выполнены условия Леммы 1 и Леммы 2. Тогда справедливо следующее выражение:

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

*Доказательство.* По Лемме 2

$$\mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})}\rho(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}},\mathbf{h},\boldsymbol{\lambda}) \leq$$

$$\leq \frac{C_l}{\min_{\mathbf{\Gamma} in \mathbb{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}\in U_{\boldsymbol{\theta}}} q_{\mathbf{w}}(\mathbf{m}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w}|\cdot|q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})-p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})|q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma}.$$

Докажем что величина

$$\iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w}|\cdot|q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})-p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})|q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma}$$

стремится к нулю. Определим случайную величину $\boldsymbol{\nu}(t), t \geq 0$ следующим образом:

$$\boldsymbol{\nu}(t) = \mathbf{max}(-t\cdot\mathbf{1},\mathbf{min}(t\cdot\mathbf{1},\mathbf{w})).$$

Данная величина совпадает с $\mathbf{w}$ при $|\mathbf{w}| < t$ и принимает значение $t$ или $-t$ при $|\mathbf{w}| \geq t$. Тогда для любого $t > 0$ справедливо:

$$\iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w}|\cdot|q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})-p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})|q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma} \leq$$

*по неравенству треугольника и используя выражение* $\mathbf{w} = \mathbf{w} + \boldsymbol{\nu}(t) - \boldsymbol{\nu}(t)$

$$\leq \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w}-\boldsymbol{\nu}(t)|\cdot|p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})-q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})|q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma}+ \qquad (2.11)$$

$$+\iint_{\mathbf{\Gamma},\mathbf{w}} |\boldsymbol{\nu}(t)|\cdot|q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})-p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})|q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma}.$$

Рассмотрим первое слагаемое суммы (2.11). Т.к. вторые моменты $\mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})}\mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}\mathbf{w}^2, \mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})}\mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})}\mathbf{w}^2$ конечны, то случайная величина $\mathbf{w}$ равномерно интегрируема как при маргинальном распределении $\int_{\mathbf{\Gamma}} q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})d\mathbf{\Gamma}$, так и при маргинальном распределении $\int_{\mathbf{\Gamma}} q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})d\mathbf{\Gamma}$. По определению равномерной интегрируемости для $\mathbf{w}$ для любого числа $\varepsilon$ существует число $t_0$, такое что для любого $t \geq t_0$, любого $\mathbf{h} \in U_{\mathbf{h}}, \boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$, справедливо выражение:

$$\mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})}\mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})}|\mathbf{w}-\boldsymbol{\nu}(t)| = \iint_{\mathbf{w},\mathbf{\Gamma}} |\mathbf{w}-\boldsymbol{\nu}(t)|q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma} \leq \varepsilon,$$

$$\mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})}\mathsf{E}_{p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})}|\mathbf{w}-\boldsymbol{\nu}(t)| = \iint_{\mathbf{w},\mathbf{\Gamma}} |\mathbf{w}-\boldsymbol{\nu}(t)|p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})d\mathbf{w}d\mathbf{\Gamma} \leq \varepsilon.$$

Тогда

$$\iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w}-\boldsymbol{\nu}(t)|\cdot|p(\mathbf{w}|\mathbf{\Gamma},\mathbf{h},\boldsymbol{\lambda})-q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})|d\mathbf{w}d\mathbf{\Gamma} \leq$$

*так как модуль разностей меньше или равен суммы модулей*

$$\iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)| p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) + \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)| q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) d\mathbf{\Gamma} d\mathbf{w} < 2\varepsilon$$

для любого $t \geq t_0$. Обозначим за $\varepsilon(t)$ минимальное число $\varepsilon$, удовлетворяющее предыдущим неравенствам. Тогда

$$\iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)| \cdot |p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})| d\mathbf{w} d\mathbf{\Gamma} \leq 2\varepsilon(t),$$

где $\lim_{t\to\infty} \varepsilon(t) = 0$.

Рассмотрим второе слагаемое.

$$\iint_{\mathbf{\Gamma},\mathbf{w}} |\boldsymbol{\nu}(t)| \cdot |q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})| d\mathbf{w} d\mathbf{\Gamma} \leq$$

*по ограниченности функции* $\boldsymbol{\nu}(t)$

$$\leq t \iint_{\mathbf{\Gamma},\mathbf{w}} |q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})| q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) d\mathbf{w} d\mathbf{\Gamma}.$$

Переходя к пределу в (2.11) получим:

$$\lim_{i\to\infty} \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w}| \cdot |q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) - p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}[i], \boldsymbol{\lambda})| q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}[i]) d\mathbf{w} d\mathbf{\Gamma} =$$

*добавим предел по* $t$*, от которого не зависит данное выражение*

$$= \lim_{t\to\infty} \lim_{i\to\infty} \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w}| \cdot |q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}[i]) - p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}[i], \boldsymbol{\lambda})| q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}[i]) d\mathbf{w} d\mathbf{\Gamma} \leq$$

*из выше написанных неравенств*

$$\lim_{t\to\infty} \lim_{i\to\infty} \iint_{\mathbf{\Gamma},\mathbf{w}} |\mathbf{w} - \boldsymbol{\nu}(t)| \cdot |p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}[i], \boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}[i])| d\mathbf{w} d\mathbf{\Gamma} +$$

$$+ \iint_{\mathbf{\Gamma},\mathbf{w}} |\boldsymbol{\nu}(t)| \cdot |q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}[i]) - p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}[i], \boldsymbol{\lambda})| q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}[i]) d\mathbf{w} d\mathbf{\Gamma} \leq$$

$$\lim_{t\to\infty} 2\varepsilon(t) + \lim_{t\to\infty} \lim_{i\to\infty} t \iint_{\mathbf{\Gamma},\mathbf{w}} |q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}[i]) - p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}_i, \boldsymbol{\lambda}) q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}[i]) = 0.$$

Последнее равенство следует из Леммы 2. Таким образом выражение

$$\left| \int_{\mathbf{\Gamma}} \frac{q_{\mathbf{w}}(\mathrm{mode} q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}{q_{\mathbf{w}}(\mathrm{mode} p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})} q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}) d\mathbf{\Gamma} \right|$$

стремится к единице, что и требовалось доказать. □

Теорема утверждает, что при устремлении параметрической сложности модели к нулю, все параметры $\mathbf{w}$ модели подлежат удалению в среднем по всем возможным значениям структуры $\mathbf{\Gamma}$ модели. Заметим, что теорема применима для случая, когда последовательность вариационных распределений $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})$ не имеет предела. Так, в случае, если структура $\mathbf{\Gamma}$ определена однозначно, последовательность $\boldsymbol{\theta}_i$ может являться последовательностью нормальных распределений, чье матожидание стремится к нулю:

$$\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\mu}_q[i], \mathbf{A}_q^{-1}[i]), \boldsymbol{\mu}_q[i] \to \mathbf{0}.$$

Априорным распределением $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) = p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ при этом может являться семейство нормальных распределений с нулевым средним:

$$p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) = \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}).$$

При этом сама последовательность распределений $\boldsymbol{\theta}[i]$ не обязана иметь предел.

## 2.3. Обобщающая задача

В данном разделе проводится анализ основных критериев выбора моделей, а также предлагается их обобщение на случай моделей, испольюзующих вариационное распределение $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})$ для аппроксимации неизвестного апостериорного распределения параметров $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$.

Рассмотрим основные статистические критерии выбора вероятностных моделей.

1. Критерий максимального правдоподобия:

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

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

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})} \log\ p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}). \tag{2.12}$$

Данное обобщение (2.12) эквивалентно критерию максимального правдоподобия при выборе в качестве $q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta})$ эмпирического распределения параметров TODO и структуры. Метод не предполагает оптимизации гиперпараметров $\mathbf{h}$. Для формального соответствия данной задачи задаче выбора модели (1.5),(1.6), т.е. двухуровневой задачи оптимизации, положим $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ :

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

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

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:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) = \tag{2.13}$$

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

Т.к. в рамках данной задачи (2.13) 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:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) = p'|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) \tag{2.14}$$

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

4. Критерий Акаике:

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

Т.к. все 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:

$$L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) = \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - \tag{2.15}$$

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

где

$$\mathbf{h} = \arg\min_{\mathbf{h}'\in U_{\mathbf{h}}} D_{\text{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big), \tag{2.16}$$

$\lambda_{\text{prune}}$ — метапараметр алгоритма, $U_{\mathbf{h}} \subset \mathbb{H}$ — область определения задачи по гиперпараметрам. Предложенное обобщение (2.15) применимо только в случае, если выражение (2.16) определено однозначно, т.е. существует единственный вектор гиперпараметров $\mathbf{h} \in U_{\mathbf{h}}$, доставляющий минимум дивергенции $D_{\text{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big)$.

5. Информационный критерий Шварца:

$$\mathrm{BIC} = \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - 0.5 \log m|\mathbb{W}|.$$

Переформулируем данный критерий аналогично критерию AIC:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \tag{2.17}$$

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

метапараметр $\lambda_{\mathrm{prune}}$ определен аналогично (2.16).

6. Метод вариационной оценки обоснованности:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \tag{2.18}$$

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

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

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

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:

$$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}_{\mathrm{train}}|\mathbf{X}_{\mathrm{train}}, \mathbf{w}, \boldsymbol{\Gamma}) + \log p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) \to \max_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}}, \tag{2.19}$$

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

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

Каждый из рассмотренных критериев удовлетворяет хотя бы одному из перечисленных свойств:

1) модель, оптимизируемая согласно критерию, доставляет максимум правдоподобия выборки;
2) модель, оптимизируемая согласно критерию, доставляет максимум оценки обоснованности;
3) для моделей, доставляющих сопоставимые значения правдоподобия выборки, выбирается модель с меньшим количеством информативных параметров.
4) критерий позволяет производить перебор структур для отбора наилучших.

Формализуем рассмотренные критерии. Оптимизационную задачу, которая удовлетворяет всем перечисленным свойствам при некоторых значинях метапараметров, будет называть *обобщающей.*

**Определение 23.** 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 (1.6)

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

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 нерперывности: 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 $p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ значимо высока. При этом соответствующие вариационные распределения $q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ 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 $q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ различаются значимо. Последний критерий говорит о том, что обобщающая задача должна позволять производить переход между различными методами выбора параметров и структуры модели непрерывно.

**Теорема 3.** Рассмотренные задачи (2.12),(2.13),(2.14),(2.15),(2.17),(2.19) не являются обобщающими.

*Доказательство.* Задачи (2.12),(2.13),(2.14),(2.15),(2.17) не имеют гиперпараметров $\mathbf{h}$, подлежащих оптимизации, поэтому не могут приближать вариационную оценку.

При использовании валидации на отложенной выборки (2.19) в функцию валидации $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ не входит ни один метапараметр, поэтому критерий перебора структур 6 для нее также не выполняется. □

Докажем также, что задача (2.18) также не является обобщающей.

**Теорема 4.** 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 нулевым не более чем счетное количество раз.
2. Выражение $\nabla_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}} q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) \log p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ ограничено на $U_{\boldsymbol{\theta}}$ некоторой случайной величиной с конечным первым моментом.

Тогда задача (2.18) не является обобщающей.

*Доказательство.* Пусть выполнены условия критерия 6 о переборе структур, и $\mathbf{h}_1, \mathbf{h}_2$ — локальные оптимумы функции $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ при метапараметрах $\boldsymbol{\lambda}$. По условию критерия соответствтие $\boldsymbol{\theta}_1 = \boldsymbol{\theta}^*(\mathbf{h}_1)$ и $\boldsymbol{\theta}_2 = \boldsymbol{\theta}^*(\mathbf{h}_2)$ должны сохраняться, т.е. для некоторого $\boldsymbol{\lambda}'$ решение нижней задачи оптимизации $\boldsymbol{\theta}^*(\mathbf{h}_1)$ должно совпадать с решением $\boldsymbol{\theta}^*(\mathbf{h}_1)$ при метапараметрах $\boldsymbol{\lambda}$. Тогда

$$0 = \nabla_{\boldsymbol{\theta}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \nabla_{\boldsymbol{\theta}} \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)|p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})) =$$

$$= \nabla_{\boldsymbol{\theta}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \nabla_{\boldsymbol{\theta}} \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)|p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda}')).$$

Сокращая равные слагаемые в равенстве получим:

$$\nabla_{\boldsymbol{\theta}} \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)|p(\boldsymbol{\Gamma}|\boldsymbol{\lambda})) = \nabla_{\boldsymbol{\theta}} \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)|p(\boldsymbol{\Gamma}|\boldsymbol{\lambda}')),$$

Из второго условия теоремы следует, что по теореме Лебега о мажорируемой сходимости осуществим переход дифференцирования под знак интеграла:

$$\int_{\mathbf{\Gamma}\in\mathbb{\Gamma}} \nabla_{\boldsymbol{\theta}_\Gamma} q(\mathbf{\Gamma}|\boldsymbol{\theta}_1)(\log\, p(\mathbf{\Gamma}|\boldsymbol{\lambda}) - \log\, p(\mathbf{\Gamma}|\boldsymbol{\lambda}'))d\mathbf{\Gamma} = 0.$$

Т.к. выражение $\nabla_{\boldsymbol{\theta}_\Gamma} q(\mathbf{\Gamma}|\boldsymbol{\theta}_1)$ принимает нулевое значение в счетном количестве точек, то выражение $\log\, p(\mathbf{\Gamma}|\boldsymbol{\lambda}) - \log\, p(\mathbf{\Gamma}|\boldsymbol{\lambda}')$ равно нулю почти всюду, что означает что метапараметр температуры $\lambda_{\text{temp}}$ равен при разных значениях метапараметров:

$$\lambda_{\text{temp}} = \lambda'_{\text{temp}}, \quad \lambda_{\text{temp}} \in \boldsymbol{\lambda}, \lambda'_{\text{temp}} \in \boldsymbol{\lambda}'.$$

Таким образом, метапараметры $\boldsymbol{\lambda}, \boldsymbol{\lambda}'$ отличаются лишь на метапараметры $\lambda_1, \lambda_2$ регуляризации ковариационной матрицы $\mathbf{A}^{-1}$. Возьмем в качестве векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2$ гиперпараметры, отличающиеся только параметрами распределения структуры:

$$\mathbf{h}_1 = [\mathbf{s}_1, \text{diag}(\mathbf{A}_1)], \mathbf{h}_2 = [\mathbf{s}_2, \text{diag}(\mathbf{A}_2)], \quad \mathbf{s}_1 \neq \mathbf{s}_2, \mathbf{A}_1 = \mathbf{A}_2.$$

Метапараметры $\lambda_1, \lambda_2$ не влияют на значение функции $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ при гиперпараметрах, отличающихся только параметрами распределения структуры, поэтому значение функции $Q$ для них будет неизменно при любых значениях $\lambda_1, \lambda_2$. Приходим к противоречию: значение $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ не меняется при изменении метапараметров $\boldsymbol{\lambda}$.

□

В качестве обобщающей задачи оптимизации предлагается оптимизационную задачу следующего вида:

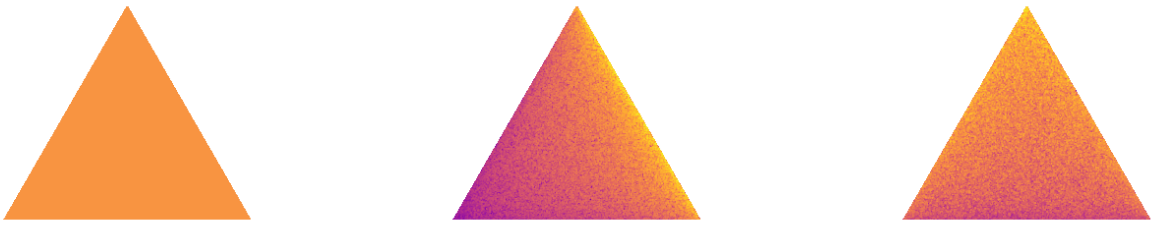
$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \tag{2.20}$$

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

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \tag{2.21}$$

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

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



(а) (б) (в)

Рис. 2.5. Пример зависимости функции $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ от гиперпараметра $\mathbf{s}$ при различных значениях метапараметров $\boldsymbol{\lambda}^{\text{Q}}_{\text{struct}}$. Темные точки на графике соответсвуют наименее предпочтительным значениям гиперпараметра. а) $\boldsymbol{\lambda}^{\text{Q}}_{\text{struct}} = [0, 0]$, б) $\boldsymbol{\lambda}^{\text{Q}}_{\text{struct}} = [1, 0]$, в) $\boldsymbol{\lambda}^{\text{Q}}_{\text{struct}} = [1, 1]$.

**Пример 3.** Рассмотрим вырожденный случай поведения функции $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$, когда $\lambda^{\text{Q}}_{\text{likelihood}} = \lambda^{\text{Q}}_{\text{prior}} = 0$. Пусть модель использует один структурный параметр, в качестве априорного распределения на структуре задано распределение Gumbel-Softmax с $\lambda_{\text{temp}}$. Пусть в качестве множества распределений $\mathfrak{P}$ используется два распределения Gumbel-Softmax, сконцентрированных близко к вершинам симплекса:

$$\mathfrak{P} = [\mathcal{GS}([0.95, 0.05, 0.05]^{\text{T}}, 1.0), \mathcal{GS}([0.95, 0.05, 0.05]^{\text{T}}, 1.0)].$$

Из определения распределения Gumbel-Softmax следует, что достаточно рассмотреть только значения параметра $\mathbf{s}$ ,находящиеся внутри симплекса. На рис. 2.5 изображены значения функции Q в зависимости от метапараметров $\boldsymbol{\lambda}^{\text{struct}}_{\text{Q}}$ и значений гиперпараметра $\mathbf{s}$ распределения на структуре. Видно, что варьируя коэффициенты метапараметров значение функции $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ значительно меняется вблизи вершин симплекса. Таким образом получается последовательность оптимизаций, схожая с полным перебором структуры.

Следующая теорема анализирует достаточные условия того, что предложенная задача оптимизации (2.20) является обобщающей.

**Теорема 5.** Пусть

1. Задано непустое множество непрерывных по параметрам распределений на структуре $\mathfrak{P}$, чьи плотности 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 $p_1 = p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$.
2. Вариационное распределение $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$ является абсолютно непрерывным, плотность которого непрерывна по метапараметрам $\boldsymbol{\lambda}$ и не принимает нулевое значение.
3. Задан компакт $U = U_{\boldsymbol{\theta}_{\mathbf{w}}} \times U_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}} \times U_{\mathbf{h}} \times U_{\boldsymbol{\lambda}}$, где параметры распределений $p \in \mathfrak{P}$ принадлежат множеству метапараметров $\boldsymbol{\lambda}$.

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 $m_\lambda$ — максимальное значение $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$ перед распределением $p_1$ из первого условия теоремы.

Тогда задача (2.20) является обобщающей на $U$.

*Доказательство.* Для доказательсta теоремы требуется доказать критерии 1-7 из определения обобщающей задачи. Выполнение критериев 1 и 2 следует из условий задачи.

Докажем критерий 3. Пусть $\lambda^{\mathrm{Q}}_{\mathrm{prior}} = 0, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$. Пусть $\lambda_1, \lambda_2, \lambda^{\mathrm{Q}}_{\mathrm{likelihood}}$ удовлетворяют седьмому условияю теоремы. Возьмем в качестве $K_1$ следующее выражение:

$$K_1 = \max_{\mathbf{h} \in U_{\mathbf{h}}} \log p(\mathbf{h}|\boldsymbol{\lambda}) - \min_{\mathbf{h} \in U_{\mathbf{h}}} \log p(\mathbf{h}|\boldsymbol{\lambda}).$$

Пусть $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}$ — гиперпараметры, удовлетворяющие условию третьего критерия:

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

. Тогда

$$Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}^*(\mathbf{h}_1))} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) -$$

$$-\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}^*(\mathbf{h}_2))} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) + \log p(\mathbf{h}_1|\boldsymbol{\lambda}) - \log p(\mathbf{h}_2|\boldsymbol{\lambda}) > K_1.$$

Отсюда следует выполнение критерия 3:

$$\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) > 0.$$

Т.к. $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} > 0$ :

$$\mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) > 0.$$

Докажем критерий 4. Пусть $\boldsymbol{\lambda}$ удовлетворяют восьмому условию теоремы и $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 0, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$. Пусть

$$K_2 = \max_{\mathbf{h} \in U_{\mathbf{h}}} \frac{1}{\lambda^{\mathrm{Q}}_{\mathrm{prior}}} \log p(\mathbf{h}|\boldsymbol{\lambda}) - \frac{1}{\lambda^{\mathrm{Q}}_{\mathrm{prior}}} \min_{\mathbf{h} \in U_{\mathbf{h}}} \log p(\mathbf{h}|\boldsymbol{\lambda}) +$$

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

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

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

Пусть $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}, Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) > K_2$. Рассмотрим разность параметрических сложностей двух векторов:

$$C_p(\boldsymbol{\theta}_2) - C_p(\boldsymbol{\theta}_1) = \min_{\mathbf{h} \in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) -$$

$$- \min_{\mathbf{h} \in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) \geq$$

*оценим снизу, а также добавим и вычтем* $D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\big)$

$$\geq \min_{\mathbf{h} \in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})\big) +$$

$$+ D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\big) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\big) =$$

*сведем выражение до* $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$

$$= Q(\mathbf{h}_1|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) - \frac{1}{\lambda^{\mathrm{Q}}_{\mathrm{prior}}} \log p(\mathbf{h}_1|\boldsymbol{\lambda}) + \frac{1}{\lambda^{\mathrm{Q}}_{\mathrm{prior}}} \log p(\mathbf{h}_2|\boldsymbol{\lambda}) +$$

$$+ \min_{\mathbf{h}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\big) >$$

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

$$> K_2 - \frac{1}{\lambda^{\mathrm{Q}}_{\mathrm{prior}}} \log p(\mathbf{h}_1|\boldsymbol{\lambda}) + \frac{1}{\lambda^{\mathrm{Q}}_{\mathrm{prior}}} \log p(\mathbf{h}_2|\boldsymbol{\lambda}) + \min_{\mathbf{h}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big)$$

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

Рассмотрим разность:

$$\min_{\mathbf{h}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\big) =$$

*т.к.* $\boldsymbol{\theta}_2$ *— решение нижней задачи оптимизации:*

$$\min_{\mathbf{h}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) - \frac{1}{\lambda^{\mathrm{L}}_{\mathrm{prior}}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) +$$

$$\max_{\boldsymbol{\theta}} \Big(\frac{1}{\lambda^{\mathrm{L}}_{\mathrm{prior}}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{h}_2, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big)\Big) \geq$$

*получим оценку снизу:*

$$\geq \min_{\mathbf{h}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) - \max_{\boldsymbol{\theta}} \frac{1}{\lambda^{\mathrm{L}}_{\mathrm{prior}}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) +$$

$$\max_{\boldsymbol{\theta}} \left( \min_{\boldsymbol{\theta}'} \frac{1}{\lambda^{\mathrm{L}}_{\mathrm{prior}}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}')} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{h}_2, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) \right) \geq$$

*оценим первое слагаемое*

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

$$\min_{\boldsymbol{\theta}} \frac{1}{\lambda^{\mathrm{L}}_{\mathrm{prior}}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - \min_{\boldsymbol{\theta}} D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{h}_2,\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) \geq$$

*оценим последнее слагаемое*

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

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

Складывая полученную оценку с $K_2 - \log \frac{1}{\lambda^{\mathrm{Q}}_{\mathrm{prior}}} p(\mathbf{h}_2|\boldsymbol{\lambda}) + \log \frac{1}{\lambda^{\mathrm{Q}}_{\mathrm{prior}}} p(\mathbf{h}_2|\boldsymbol{\lambda})$ получаем разность параметрических сложностей больше нуля, что и требовалось доказать.

Докажем критерий 5. 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 функции $L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$ и $Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda})$ можно записать как:

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

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

Двухуровневая задача оптимизации совпадает с оптимизацией вариационной оценки обоснованности, что и требовалось доказать.

Докажем критерий 6. Пусть задан вектор метапараметров $\boldsymbol{\lambda}$, удовлетворяющий девятому условию теоремы и $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$. По условию теоремы во множество $\mathfrak{P}$ входит хотя бы одно распределение Gumbel-Softmax:

$$p_1 \sim \mathcal{GS}, p \in \mathfrak{P}.$$

Возьмем в качестве $K_4$ следующее выражение:

$$K_4 = \frac{\max_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) - \min_{\mathbf{h}} Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda})}{m_\lambda},$$

где $m_\lambda$ — максимальное значение коэффициента $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$ перед $p_1$. Пусть заданы векторы гиперпараметров $\mathbf{h}_1, \mathbf{h}_2$, такие что $Q(\mathbf{h}_1|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) - Q(\mathbf{h}_2|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) > 0$ и

$$D_{\mathrm{KL}}\big(p(\mathbf{h}_1|\boldsymbol{\lambda})||p(\mathbf{h}_2,\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) > K_4$$

,

$$D_{\mathrm{KL}}\big(p(\mathbf{h}_2,\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})||p(\mathbf{h}_1|\boldsymbol{\lambda})\big) > K_4.$$

Пусть вектор метапараметров $\boldsymbol{\lambda}'$ отличается от $\boldsymbol{\lambda}$ лишь метапараметром $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$. Для обоих векторов метапараметров нижняя задача оптимизации $L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$ совпадает, поэтому выполняется первое условие критерия.

Положим для $\boldsymbol{\lambda}'$ метапараметр $\lambda^{Q}_{\mathrm{struct}} \in \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$ перед распределением $p_1$ равным максимальному значению $m_\lambda$. Положим также значение параметров данного распределения равным параметрам распределения $p(\mathbf{h}_1,\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})$ :

$$p_1 = p(\mathbf{h}_1,\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda}).$$

Для остальных распределений $p' \in \mathfrak{P}$ положим коэффициент $\lambda^{Q}_{\mathrm{struct}} \in \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$ равным нулю. Тогда справедливо следующее неравенство:

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

$$= Q(\mathbf{h}_2|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) - Q(\mathbf{h}_1|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) + m_\lambda \lambda^{Q}_{\mathrm{struct}} D_{\mathrm{KL}}\big(p(\mathbf{h}_2,\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})||p(\mathbf{h}_1,\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) =$$
$$= Q(\mathbf{h}_2|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) - Q(\mathbf{h}_1|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) + m_\lambda K_4 > 0.$$

что и требовалось доказать.

Докажем критерий 7. Достаточным условием непрерывности функций $L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$, $Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda})$ является непрерывность входящих в нее слагаемых.

Слагаемое $\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma})$ не зависит от метапараметров $\boldsymbol{\lambda}$. Слагаемое $\log p(\mathbf{h}|\boldsymbol{\lambda})$ непрерывно по метапараметрам по свойству обратного гамма-распределения.

Достаточным условием непрерывности функций вида $D_{\mathrm{KL}}\big(p_1||p_2\big)$ является непрерывность по метапараметрам функций $p_1(\log p_1 - \log p_2)$ почти всюду и ограниченность интегрируемой функцией. Априорные распределения задаются нерперывными функциями плотности $p(\mathbf{w}|\boldsymbol{\Gamma},\mathbf{h},\boldsymbol{\lambda}), p(\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})$, не принимающими нулевое значение, и являющимися непрерывными по метапараметрам. Функция $q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})$ принимает нулевое значение лишь в конечном количество точек, поэтому функция $q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})(\log q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}) - \log p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda}))$ почти всюду непрерывна по метапараметрам. Она ограничена на компакте $U_{\boldsymbol{\lambda}}$, поэтому слагаемое $D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big)$ является непрерывным по метапараметрам. Выражения вида $p(\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})(\log p(\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda}) - \log p), p \in \mathfrak{P}$ также являются непрерывными по метапараметрам и ограниченными, поэтому слагаемые вида $D_{\mathrm{KL}}\big(p(\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})||p\big)$ являются непрерывными. Поэтому функции $L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda}), Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda})$ являются непрерывными по метапараметрам, что и требовалось доказать. □

Метапараметрами данной задачи (2.20) являются коэффициенты $\lambda^{\mathrm{L}}_{\mathrm{prior}}, \lambda^{\mathrm{Q}}_{\mathrm{prior}}$, отвечающие за регуляризацию верхней и нижней задачи оптимизации, коэффициент $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}$ отвечает за максимизацию правдоподобия, а также параметры распрделений $\mathfrak{P}$ и вектор коэффициентов перед ними $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$.

Условия 7-9 теоремы задают вид области $U$, на которой представленная оптимизационная задача является обобщающей. Условие 7 выполняется при небольшом разбросе значений $\log p(\mathbf{h}|\boldsymbol{\lambda})$ в зависимости от $\lambda_1, \lambda_2$. Т.к. эти метапараметры выполняют роль регуляризатора, для области гиперпараметров $U_{\mathbf{h}}$, выбранной адекватно, данное условие выполняется.

В случае, если $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ — нормальное распределение, а $q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ — распределение Gumbel-softmax, такие что для любого $\mathbf{h} \in U_{\mathbf{h}}$ существует $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$:

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

а также полагая что $\log p(\mathbf{h}|\boldsymbol{\lambda})$ приблизительно равен для всех $\mathbf{h} \in U_{\mathbf{h}}$, восьмое условие можно представить в следующем виде:

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

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

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

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

Данное условие требует существования набора метапараметров $\boldsymbol{\lambda}$, такого что максимальная разница дивергенций на $U$ больше, чем максимальная разница между усредненными по $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$ логарифмами правдоподобия выборки, поделенными на $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}$. Условие будет выполняться при достаточно больших $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}$. Условие 9 выполняется при достаточно больших значениях метапараметра $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$.

## 2.4. Анализ обобщающей задачи

В данном разделе рассматриваются свойства предложенной задачи при различных значениях метапараметров, а также характер ассимптотического поведения задач. Следующие теоремы говорят о соответствии предлагаемой обобщающей задачи вероятностной модели. В частности, задача оптимизации параметров и гиперпараметров соответствует двухуровневому байесовскому выводу.

**Теорема 6.** 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 (2.20) 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.

*Доказательство.* Так как параметры $\boldsymbol{\theta}$ не зависят от слагаемых при коэффициентах $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$, а также от $\log p(\mathbf{h}|\boldsymbol{\lambda})$, то при $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = \lambda^{\mathrm{L}}_{\mathrm{prior}} = 1$ как верхняя, так и нижняя задачи оптимизации (2.20) эквивалентны оптимизации вариационной оценки обоснованности, поэтому первое утверждение выполняется.

Докажем второе утверждение. Рассмотрим оценку обоснованности модели:

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

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

Из данного равенства следует:

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

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

где правая часть равенства соответствует вариационной оценки обоснованности. Выражение $\log p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ не зависит от вариационного распределения $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$, поэтому максимизации вариационной оценки эквивалентна минимизации дивергенции $D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})\big)$.

Докажем третье утверждение. Т.к. вариационное распределение $q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$ декомпозируется на $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}), q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$, апостериорное распределение $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ декомпозируется на $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}), p(\boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$, поэтому достижимо значение нулевое значение дивергенции: $D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})\big) = 0$. Она представима в виде (2.8). Отсюда следует что соотстветсвующие вариационные и апостериорные распределения совпадают. □

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

**Теорема 7.** 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 совопкупности.

*Доказательство.* Рассмотрим величину $\frac{1}{m} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$:

$$\frac{1}{m} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \frac{1}{m} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \tag{2.22}$$

$$- \frac{\lambda^{\mathrm{L}}_{\mathrm{prior}}}{m} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big).$$

При $m \gg 0$ по усиленному закону больших чисел данная функция может быть аппроксимирована следующим образом:

$$\frac{1}{m} L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) \approx \mathsf{E}_{y,\mathbf{x}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma})$$

$$- \frac{\lambda^{\mathrm{L}}_{\mathrm{prior}}}{m} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big).$$

Аналогично рассмотрим вариационную оценку обоснованности для произвольной выборки мощностью $m_0 = \frac{m}{\lambda^{\mathrm{L}}_{\mathrm{prior}}}$, усредненную на мощность выборки:

$$\frac{1}{m_0} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \frac{1}{m_0} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) \approx \tag{2.23}$$

$$\approx \mathsf{E}_{y,\mathbf{x}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \frac{1}{m_0} D_{\mathrm{KL}}\big(p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})||q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})\big) =$$

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

Таким образом, задачи оптимизации функций (2.22),(2.23) совпадают, что и требовалось доказать. □

Теорема показывает, что для достаточно большого $m$ и $\lambda^{\mathrm{L}}_{\mathrm{prior}} > 0, \lambda^{\mathrm{L}}_{\mathrm{prior}} \neq 1$ оптимизация параметров и гиперпараметров эквивалентна нахождению оценки обоснованности для выборки другой мощности: чем выше значение $\lambda^{\mathrm{L}}_{\mathrm{prior}}$, тем выше мощность выборки, для которой проводится оптимизация.

Таким образом, предлагаемая обобщающая задача производит оптимизацию вариационной оценки обоснованности с различными эффекитвными размерами

выборок. Чем больше размер выборки, тем больше влияние априорного распределения, которое выступает в качестве регуляризатора. Сложность модели назначается следующим образом:

1. варьированием сложности на верхнем уровне оптимизации оптимизации с использованием коэффициента $\lambda^{\mathrm{Q}}_{\mathrm{prior}}$;
2. варьированием сложности на нижнем уровне оптимизации оптимизации с использованием коэффициента $\lambda^{\mathrm{L}}_{\mathrm{prior}}$;
3. варьированием сложности на обоих уровнях оптимизации.

Рассмотрим различие вариантов 1-3 на примере.

**Пример 4.** Назначим $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = 0$. Требуется уменьшить вклад априорного распределения в итоговую оптимизацию. При варьировании нижней задачи оптимизации ($\lambda^{\mathrm{L}}_{\mathrm{prior}} \to 0$) оптимизационная задача становится эквивалента методу максимального правдоподобия:

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

При этом верхняя задача $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \to \max_{\mathbf{h}}$ не имеет смысла, т.к. параметры $\boldsymbol{\theta}$ не зависят от гиперпараметров $\mathbf{h}$.

При варьировании только верхней задачи оптимизации ($\lambda^{\mathrm{Q}}_{\mathrm{prior}} \to 0, \lambda^{\mathrm{L}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 1$), на нижнем уровне задача $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ совпадает с задачей выбора обоснованных параметров при фиксированном значении гиперпараметров $\mathbf{h}$:

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

При этом на верхнем уровне оптимизации выбираются гиперпараметры $\mathbf{h}$, при которых параметры будут доставать максимум правдоподобия с точностью до регуляризации:

$$Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \to_{\lambda^{\mathrm{Q}}_{\mathrm{prior}} \to 0} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) + \log p(\mathbf{h}|\boldsymbol{\lambda}).$$

Данный пример показывает, что при варьировани сложности на разлинчных уровнях оптимизации приводит к значительно различающимся результатам: снижение значения коэффициента $\lambda^{\mathrm{Q}}_{\mathrm{prior}}$ на верхнем уровне оптимизации приводит к выбору модели с параметрами, соответствующими максимуму вариационной оценки при гиперпараметрах, выбранных согласно критерию максимального правдоподобия. Варьирование сложности на нижнем уровне оптимизации приводит вид всей оптимизации к критерию максимального правдоподобия.

Докажем теорему об оценке разности параметрических сложностей.

**Лемма 3.** Пусть задан компакт $U = U_{\mathbf{h}} \times U_{\boldsymbol{\theta}}$ и $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$. Пусть также решение задачи

$$\min_{\mathbf{h} \in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) \tag{2.24}$$

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 $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}, \lambda^{\mathrm{L}}_{\mathrm{prior}}, \lambda_{\mathrm{temp}}, \lambda_1, \lambda_2$.

Тогда справедливо следующее неравенство:

$$D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda}')\big) < D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big),$$

где $\mathbf{h}_1, \boldsymbol{\theta}_1$, $\mathbf{h}_2, \boldsymbol{\theta}_2$ — решения задачи (2.20) при $\lambda^{\mathrm{Q}}_{\mathrm{prior}_1}, \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$,

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}^*(\mathbf{h}_1), \quad \boldsymbol{\theta}_2 = \boldsymbol{\theta}^*(\mathbf{h}_2),$$

$\boldsymbol{\lambda}'$ — вектор метапараметров, содержащий метапараметры $\lambda_{\mathrm{temp}}, \lambda_1, \lambda_2$

*Доказательство.* Заметим, что выражение вида $D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{h}_1, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big)$ зависит только от метапараметров $\boldsymbol{\lambda}' = [\lambda_{\mathrm{temp}}, \lambda_1, \lambda_2]$ и не зависит от $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}, \lambda^{\mathrm{L}}_{\mathrm{prior}}, \lambda^{\mathrm{Q}}_{\mathrm{prior}}, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$.

Пусть $\mathbf{h}_1, \boldsymbol{\theta}_1$, $\mathbf{h}_2, \boldsymbol{\theta}_2$ — решения задачи (2.20) при $\lambda^{\mathrm{Q}}_{\mathrm{prior}_1}, \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$. Тогда справедлисва система неравенств:

$$\begin{aligned}
&\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \\
&- \lambda^{\mathrm{Q}}_{\mathrm{prior}_1} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda}')\big) + \log p(\mathbf{h}_1|\boldsymbol{\lambda}_1) > \\
&> \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \\
&\lambda^{\mathrm{Q}}_{\mathrm{prior}_1} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big) + \log p(\mathbf{h}_2|\boldsymbol{\lambda}_2);
\end{aligned}$$

$$\begin{aligned}
&\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \\
&- \lambda^{\mathrm{Q}}_{\mathrm{prior}_2} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big) + \log p(\mathbf{h}_2|\boldsymbol{\lambda}_2) > \\
&> \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} \mathsf{E}_{q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \\
&- \lambda^{\mathrm{Q}}_{\mathrm{prior}_2} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda}')\big) + \log p(\mathbf{h}_1|\boldsymbol{\lambda}_1).
\end{aligned}$$

Складывая неравенства получим следующее выражение:

$$\begin{aligned}
&(\lambda^{\mathrm{Q}}_{\mathrm{prior}_2} - \lambda^{\mathrm{Q}}_{\mathrm{prior}_1}) D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda}')\big) > \\
&> (\lambda^{\mathrm{Q}}_{\mathrm{prior}_2} - \lambda^{\mathrm{Q}}_{\mathrm{prior}_1}) D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big).
\end{aligned}$$

Т.к. по условию $\lambda^{\mathrm{Q}}_{\mathrm{prior}_1} > \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$, то отсюда следует:

$$D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda}')\big) < D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big),$$

что и требовалось доказать. □

**Теорема 8.** Пусть

1. Ввыполнены условия Леммы 3.

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 (2.24) 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.

Тогда справедлива следующая оценка разности параметрических сложностей:

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

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

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

*Доказательство.* Положим $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$ — два набора метапараметров с фискированными значениями метапараметров, соответствующих условиями теоремы и отличающихся лишь значениями $\lambda^{\mathrm{Q}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{prior}_1}, \lambda^{\mathrm{Q}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$. Рассмотрим разность параметрических сложностей:

$$C_p(\boldsymbol{\theta}_1|U_{\mathbf{h}}, \boldsymbol{\lambda}_1) - C_p(\boldsymbol{\theta}_2|U_{\mathbf{h}}, \boldsymbol{\lambda}_2) =$$

*по определению параметрической сложности:*

$$= \min_{\mathbf{h}\in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}')\big) - \min_{\mathbf{h}\in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}')\big) <$$

*используя оценку сверху:*

$$< D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}')\big) - \min_{\mathbf{h}\in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}')\big) =$$

*добавляя и вычитая слагаемое* $D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big)$:

$$= D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big) - \min_{\mathbf{h}\in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big) +$$

$$+ D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big).$$

По лемме 3 следует:

$$D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda}')\big) - \min_{\mathbf{h}\in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big) +$$

$$+ D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big) - D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big) <$$
$$D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big) - \min_{\mathbf{h}\in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda}')\big).$$

Обозначим за $\mathbf{h}'$ — решение задачи (2.24). Тогда справедливо следующее выражение:

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

$$D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h}_2,\boldsymbol{\lambda}')\big) - D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h}',\boldsymbol{\lambda}')\big) =$$

$$\frac{1}{\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}}\big(Q(\mathbf{h}_2|\boldsymbol{\theta}_2,\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2) - Q(\mathbf{h}'|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2)\big).$$

Т.к.$Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda})$ — вогнутая, то справедливо равенство

$$Q(\mathbf{h}_2|\boldsymbol{\theta}_2,\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2) - Q(\mathbf{h}'|\boldsymbol{\theta}_2,\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}) \leq \nabla_{\mathbf{h}}(Q(\mathbf{h}_2|\boldsymbol{\theta}_2,\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2))||\mathbf{h}_2-\mathbf{h}'|| \leq$$

$$\leq \nabla_{\mathbf{h}}(Q(\mathbf{h}_2|\boldsymbol{\theta}_2,\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2)) \max_{\mathbf{h}_1,\mathbf{h}_2}||\mathbf{h}_1-\mathbf{h}_2||.$$

Рассмотрим выражение $\nabla_{\mathbf{h}}Q(\mathbf{h}_2|\boldsymbol{\theta}_2,\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2)$. Из [**?**] следует равенство:

$$\nabla_{\mathbf{h}}Q(\mathbf{h}_2|\boldsymbol{\theta}^*(\mathbf{h}_2),\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2) = \nabla_{\mathbf{h}}Q(\mathbf{h}_2|\boldsymbol{\theta}_2,\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2) -$$

$$-\nabla_{\boldsymbol{\theta},\mathbf{h}}(L(\boldsymbol{\theta}_2|\mathbf{y},\mathbf{X},\mathbf{h}_2,\boldsymbol{\lambda}_2))^{\mathsf{T}}\nabla^2_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}_2|\mathbf{y},\mathbf{X},\mathbf{h}_2,\boldsymbol{\lambda}_2))^{-1}\nabla_{\boldsymbol{\theta}}Q(\mathbf{h}_2|\boldsymbol{\theta}_2,\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}),$$

где левая часть равенства градиент от $Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda})$ как от сложной функции, где $\boldsymbol{\theta}^*$ — решение нижней задачи оптимизации. Т.к. $\mathbf{h}_2$ — решение задачи оптимизации (**??**), то $\nabla_{\mathbf{h}}Q(\mathbf{h}_2|\boldsymbol{\theta}^*(\mathbf{h}_2),\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2) = 0$. Отсюда следует:

$$Q(\mathbf{h}_2|\boldsymbol{\theta}_2,\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2) - Q(\mathbf{h}'|\boldsymbol{\theta}_2,\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2) \leq$$

$$\leq \nabla_{\boldsymbol{\theta},\mathbf{h}}(L(\boldsymbol{\theta}_2|\mathbf{y},\mathbf{X},\mathbf{h}_2,\boldsymbol{\lambda}_2))^{\mathsf{T}}\nabla^2_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}_2|\mathbf{y},\mathbf{X},\mathbf{h}_2,\boldsymbol{\lambda}_2))^{-1}\nabla_{\boldsymbol{\theta}}Q(\mathbf{h}_2|\boldsymbol{\theta}_2,\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2)\max_{\mathbf{h}_1,\mathbf{h}_2}||\mathbf{h}_1-\mathbf{h}_2$$

Функция $L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})$ состоит из двух слагаемых, одно из которых не зависит от $\mathbf{h}$, поэтому

$$\nabla_{\boldsymbol{\theta},\mathbf{h}}(L(\boldsymbol{\theta}_2|\mathbf{y},\mathbf{X},\mathbf{h}_2,\boldsymbol{\lambda}_2))^{\mathsf{T}} = \lambda^{\mathrm{L}}_{\mathrm{prior}}\nabla_{\boldsymbol{\theta},\mathbf{h}}(D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||[p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h}_2,\boldsymbol{\lambda}')\big))^{\mathsf{T}}.$$

Т.к. $\boldsymbol{\theta}_2$ — оптимум функции $L(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathbf{h}_2,\boldsymbol{\lambda}_2)$, то

$$\nabla_{\boldsymbol{\theta}}\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - \nabla_{\boldsymbol{\theta}}\lambda^{\mathrm{L}}_{\mathrm{prior}}D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h}_2,\boldsymbol{\lambda}_2)\big) = 0,$$

$$\nabla_{\boldsymbol{\theta}}Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2) = \nabla_{\boldsymbol{\theta}}\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) -$$

$$-\lambda^{\mathrm{Q}}_{\mathrm{prior}_2}\nabla_{\boldsymbol{\theta}}D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h}_2,\boldsymbol{\lambda}_2)\big) =$$

$$= (\lambda^{\mathrm{Q}}_{\mathrm{prior}_2} - \lambda^{\mathrm{L}}_{\mathrm{prior}})\nabla_{\boldsymbol{\theta}}D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h}_2,\boldsymbol{\lambda}_2)\big).$$

С учетом переписанных выражений $\nabla_{\boldsymbol{\theta},\mathbf{h}}(L(\boldsymbol{\theta}_2|\mathbf{h}_2,\mathbf{X},\mathbf{h},\boldsymbol{\lambda}))^{\mathsf{T}}$, $\nabla_{\boldsymbol{\theta}}Q(\mathbf{h}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda})$ получаем:

$$\nabla_{\mathbf{h}}Q(\mathbf{h}_2|\boldsymbol{\theta}^*(\mathbf{h}_2),\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2) = \nabla_{\mathbf{h}}Q(\mathbf{h}_2|\boldsymbol{\theta}_2,\mathbf{X},\boldsymbol{\theta},\boldsymbol{\lambda}_2) -$$

$$-\lambda^{\mathrm{L}}_{\mathrm{prior}}(\lambda^{\mathrm{Q}}_{\mathrm{prior}} - \lambda^{\mathrm{L}}_{\mathrm{prior}})\nabla_{\boldsymbol{\theta},\mathbf{h}}(D_{\mathrm{KL}}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||[p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h}_2,\boldsymbol{\lambda}_2)\big))^{\mathsf{T}}\times$$

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

Отсюда следует доказываемое неравенство. □

Оценка, полученная в данной теореме, зависит от метапараметров и гиперпараметров, использованных только в задаче оптимизации при $\lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$. Верхняя оценка разности параметрических сложностей обращается в ноль при $\lambda^{\mathrm{Q}}_{\mathrm{prior}_2} = \lambda^{\mathrm{L}}_{\mathrm{prior}}$ и при $\lambda^{\mathrm{L}}_{\mathrm{prior}} = 0$. Последний случай соответствует вырожденному случаю, когда нижняя задача оптимизации эквивалентна оптимизации правдоподобия выборки, и оценка параметрической разности параметрической сложности напрямую следует из Леммы 3.

Следующая теорема анализирует оптимизацию при $\frac{\lambda^{\mathrm{Q}}_{\mathrm{prior}}}{\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}} = \lambda^{\mathrm{L}}_{\mathrm{prior}}$. В частности, если $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 1$, то такая оптимизация соответствует оптимизации вариационной оценки обоснованности на обеих уровнях оптимизации для выборки размера $[\frac{m}{\lambda^{\mathrm{L}}_{\mathrm{prior}}}]$, о чем говорилось в Теореме 2.4.

**Теорема 9.** Пусть $\frac{\lambda^{\mathrm{Q}}_{\mathrm{prior}}}{\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}} = \lambda^{\mathrm{L}}_{\mathrm{prior}}$. Тогда задача оптимизации (2.20) представима в виде одноуровневой задачи оптимизации:

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

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

*Доказательство.* Т.к. выполнено равенство $\frac{\lambda^{\mathrm{Q}}_{\mathrm{prior}}}{\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}} = \lambda^{\mathrm{L}}_{\mathrm{prior}}$, то нижняя задача оптимизации экивалентна следующей задаче:

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

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

Параметры $\boldsymbol{\theta}$ вариационного распределения $q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})$ не зависят от слагаемых вида $\log p(\mathbf{h}|\boldsymbol{\lambda})$ и $D_{\mathrm{KL}}\big(p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})||p'\big), p' \in \mathfrak{P}$, поэтому нижняя задача оптимизации эквивалентна следующей задаче:

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

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

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

для любого вектора $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}}$.

Поэтому верхняя и нижняя задачи совпадают:

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

где

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

Из свойства

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

следует доказательство теоремы. □

Для вычисления приближенного значения функций $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ и $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ предлагается использовать приближение методом Монте-Карло с порождением $R$ реализаций величин $\mathbf{w}, \boldsymbol{\Gamma}$. Т.к. эти функции состоят из слагаемых вида $\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma})$, $D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big)$, $\log p(\mathbf{h}|\boldsymbol{\lambda})$, $D_{\mathrm{KL}}\big(p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})||p'\big), p' \in \mathfrak{P}$, то рассмотрим численные приближения каждого из этих слагаемых.

Выражение $\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma})$ предлагается вычислять следующим образом:

$$\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) \approx \frac{1}{R} \sum_{r=1}^{R} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}_r, \boldsymbol{\Gamma}_r),$$

где $\boldsymbol{\Gamma}_r$ — реализация случайной величины, полученная по форумле (2.2), $\mathbf{w}_r$ — реализация случайной величины, полученная по формуле:

$$\mathbf{w}_r = \boldsymbol{\mu}_q + \boldsymbol{\varepsilon}^{\mathsf{T}} \boldsymbol{\alpha}_q.$$

Выражение $D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big)$ декомпозируется на два слагаемых:

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

$$+ \int_{\boldsymbol{\Gamma}} \int_{\mathbf{w}} q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) \log \frac{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}{p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})} dq_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) dq_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}).$$

Для первого слагаемого предлагается использовать следующую формулу:

$$D_{\mathrm{KL}}\big(q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})||p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) \approx \frac{1}{R} \sum_{r=1}^{R} \log q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}_r|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) - \log p(\boldsymbol{\Gamma}_r|\mathbf{h}, \boldsymbol{\lambda}). \tag{2.25}$$

Для второго слагаемого справедлива следующая формула:

$$\int_{\boldsymbol{\Gamma}} \int_{\mathbf{w}} q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) \log \frac{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})}{p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})} dq_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}) dq_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) \approx$$

$$\approx \frac{1}{2R} \sum_{r=1}^{R} \sum_{(j,k) \in E} \sum_{l=1}^{K_{j,k}} \big( (\gamma_l^{j,k})^{-1} \mathrm{tr} \mathbf{A}_q^{-1} \mathbf{A} + (\gamma_l^{j,k})^{-1} \boldsymbol{\mu}_q^{\mathsf{T}} \mathbf{A}^{-1} \boldsymbol{\mu}_q + \log \frac{\gamma_l^{j,k} \det \mathbf{A}}{\det \mathbf{A}_q} \big) - \frac{1}{2} |\mathbb{W}|.$$

**Вычислительный эксперимент**

Для анализа предлагаемого метода выбора структуры модели было проведено два эксперимента. Первый эксперимент проводился на синтетических данных. Цель эксперимента — подтверждение рассмотренных в данном разделе свойств обобщающей задачи оптимизации.

**Эксперимент на синтетической выборке**

Выборка мощностью 50 объектов была порождена по следующему правилу:

$$\mathbf{x} \in \mathbb{R}^1 1, x[j] \sim \mathcal{N}(\mathbf{0}, 1), \mathbf{x} \in \mathbf{X}.$$

$$y = \tanh(x[0]) + 0.5\varepsilon, \varepsilon \sim \mathcal{N}(0, 1).$$

Рассматривалось параметрическое семейство моделей вида:

$$\mathbf{f} = \gamma_0^{0,1}\mathbf{f}_0 + \gamma_1^{0,1}\mathbf{f}_1 + \gamma_2^{0,1}\mathbf{f}_2,$$

где

$$\mathbf{f}_0 = w_0,$$
$$\mathbf{f}_1 = \tanh(w_1 x[0]),$$
$$\mathbf{f}_2 = \tanh(\mathbf{x}\mathbf{w}_2),$$

где $x[j]$ — $j$-я компонента объекта $\mathbf{x}$. Из определения данного параметрического семейства следует, что любая модель представляется в виде линейной комбинации трех подмодели: константной модели $\mathbf{f}_0$, модели $\mathbf{f}_1$ и переусложненной модели $\mathbf{f}_2$, которая использует признаки объекта, не использовавшиеся при порождении меток объектов.

Были проведены эксперименты с различными значениями метапараметров:

1. $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 0.01, \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \lambda^{\mathrm{L}}_{\mathrm{prior}} = 100, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$;
2. $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = \lambda^{\mathrm{L}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$;
3. $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 100.0, \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \lambda^{\mathrm{L}}_{\mathrm{prior}} = 0.01, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$;

Рассматриваемые значения метапараметров удовлетворяют Теореме 9, поэтому задача сводится к одноуровневой задаче оптимизации. Распределение на гиперпараметрах $p(\mathbf{h}|\boldsymbol{\lambda})$ рассматривалось как равномерное и не учитывалось в оптимизации. Эксперимент был запущен с различными метапараметрами температуры: $\lambda_{\mathrm{temp}} = 0.2, 1.0, 10.0$. Оптимизация проводилось с использованием оператора оптимизации Adam [111]. Для каждого набора метапараметров проводилось 5 оптимизаций, для каждой оптимизации проводилось 1000 итераций. На каждой итерации использовалось $r = 3$ реализации каждой случайной величины. Для улучшения сходимости задачи первые 500 итераций оптимизации проводились с использованием упрощенного априорного распределения $\mathbf{w} \sim \mathcal{N}(0, \mathbf{A})$, после чего значения диагонали матрицы $\mathbf{A}$ устанавливались равными $\exp(10.0)$, после чего оптимизация проводилась с указанным в данном разделе априорным распределением.

График распределения полученных структур при различных значениях метапараметров приведен на Рис. 2.6. На графике видно, что чем больше влияние априорного распределения (т.е. чем больше значение метапараметров $\lambda^{\mathrm{L}}_{\mathrm{prior}}$, $\lambda^{\mathrm{Q}}_{\mathrm{prior}}$), тем меньше распределение структур сконцентрировано на модели $\mathbf{f}_2$, как на модели, имеющий наибольшее количество параметров. Чем меньше влияние

априорного распределения и больше влияние значения правдоподобия выборки, тем меньше распределение структур сконцентрировано на модели $\mathbf{f}_0$, как на модели, описывающей выборку наихудшим образом. При увеличении температуры концентрация структур смещается ближе к центру.

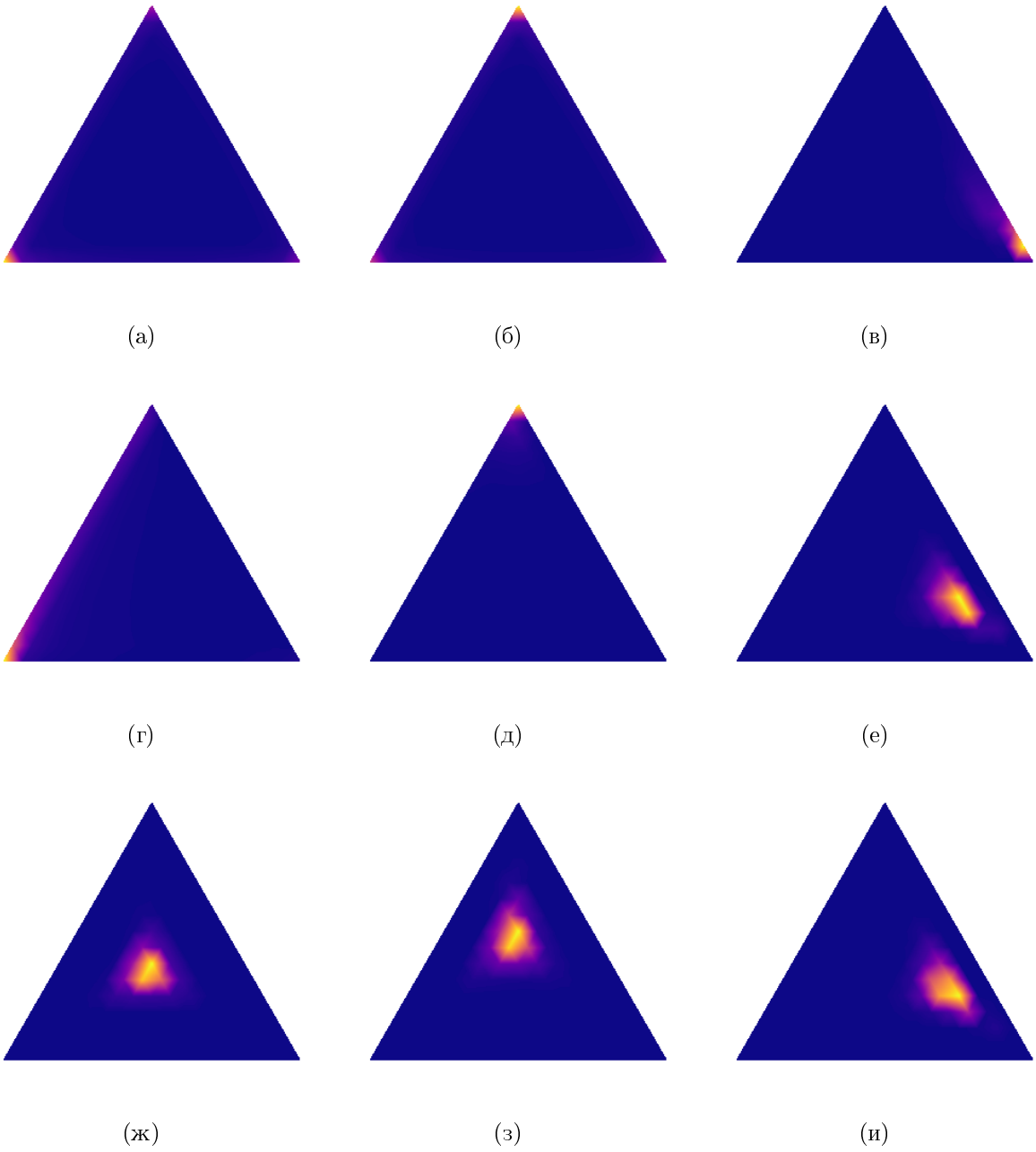
График зависимости полученных моделей от первой компоненты объекто выборки представлен на Рис. 2.7. На данном графике видно, уменьшение значения метапараметров $\lambda^{\mathrm{L}}_{\mathrm{prior}}$, $\lambda^{\mathrm{Q}}_{\mathrm{prior}}$ ведет к переобучению модели. Увеличение температуры $\lambda_{\mathrm{temp}}$ ведет к разбросу значений структуры $\mathbf{\Gamma}$ и увеличению дисперсии предсказаний модели.

График относительной плотности параметров моделей представлен на Рис. 2.8. Относительная плотность параметров соответствует наиболее вероятным по вариационному распределению структурам.

Для анализа возможности перехода между структурами был проведен эксперимент с параметрами $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 0.01, \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \lambda^{\mathrm{L}}_{\mathrm{prior}} = 100, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = [-1, -1], \lambda_{\mathrm{temp}} = 1.0$. В качестве структур $\mathfrak{P}$ выступали следующие структуры:

$$p_1 = \mathcal{GS}(0.1, [0.99, 0.05, 0.05]), \quad p_2 = \mathcal{GS}(0.1, [0.05, 0.05, 0.99]).$$

Данные структуры соответствуют распределениям структур, сконцентрированным близко к моделям $\mathbf{f}_0, \mathbf{f}_2$. Гистограмма итоговых распределений для данной задачи оптимизации представлена на Рис. 2.9. График показывает, что в отличие от оптимизации с $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$, которая представлена на Рис. 2.6д, при использовании данного слагаемого вариационное распределение $q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ сконцентрировано у модели $\mathbf{f}_1$. Заметим, что данная регуляризация напрямую влияет только на априорное распределение структур $p(\mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$, а не на вариационное распределение $q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$, поэтому итоговое распределение $q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ изменяется значительно только при равенстве (???) суммы остальных слагаемых обобщающей задачи оптимизации.

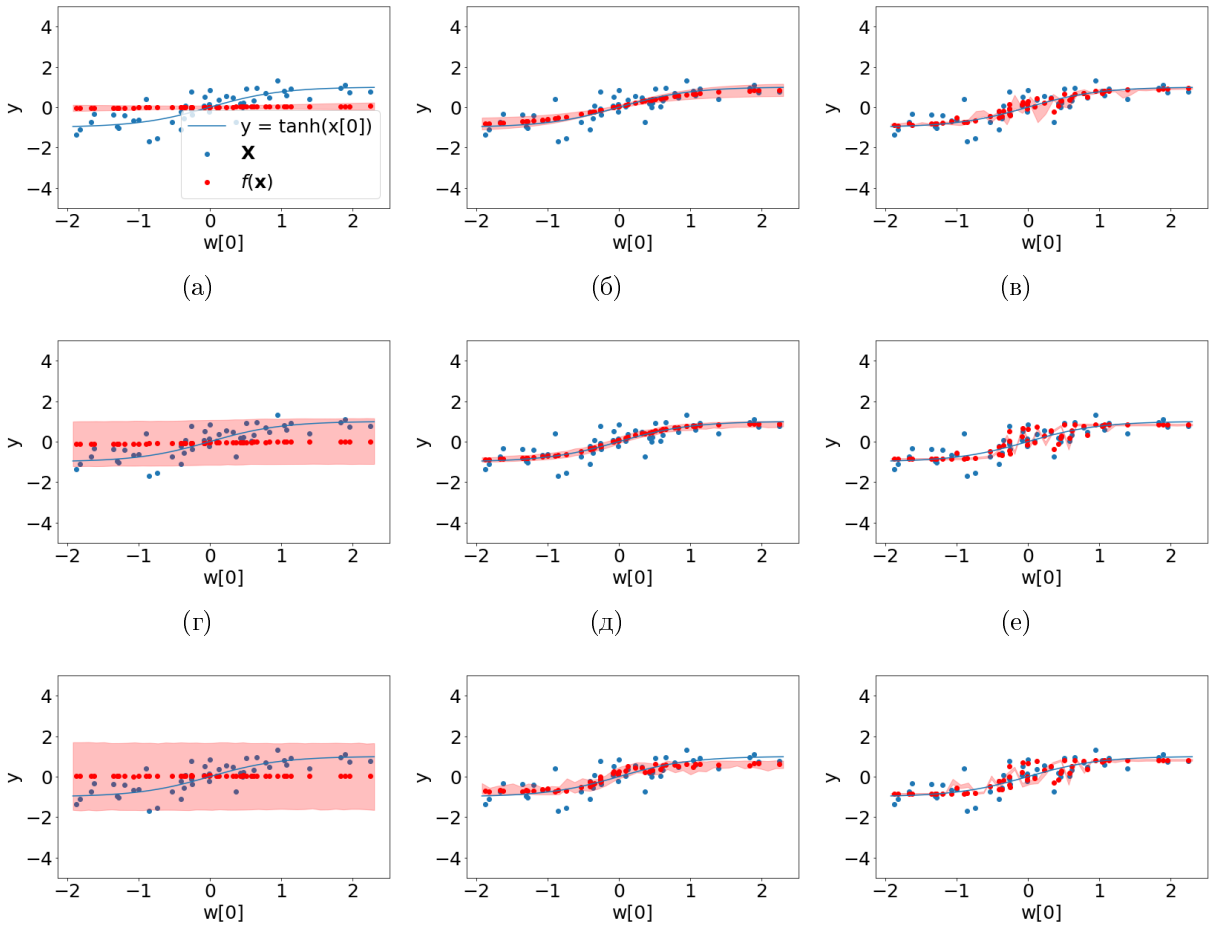


(а) (б) (в)

(г) (д) (е)

(ж) (з) (и)

Рис. 2.6. Гистограмма итогового вариационного распределения $q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}})$ структур при различных значениях метапараметров. Левый нижний угол соответствует модели $\mathbf{f}_0$, правый нижний угол соответствет модели $\mathbf{f}_1$, верхний угол соответствует модели $\mathbf{f}_2$. Первый столбец: $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 0.01, \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \lambda^{\mathrm{L}}_{\mathrm{prior}} = 100, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$, второй столбец: $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = \lambda^{\mathrm{L}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$, третий столбец: $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 100.0, \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \lambda^{\mathrm{L}}_{\mathrm{prior}} = 0.01, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$. Первая строка: $\lambda_{\mathrm{temp}} = 0.2$, вторая строка: $\lambda_{\mathrm{temp}} = 1.0$, третья строка: $\lambda_{\mathrm{temp}} = 10.0$.

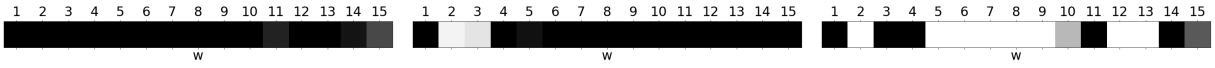



(а) (б) (в)

(г) (д) (е)

(ж) (з) (и)

Рис. 2.7. График итоговых моделей. Первый столбец: $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 0.01, \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \lambda^{\mathrm{L}}_{\mathrm{prior}} = 100, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$, второй столбец: $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = \lambda^{\mathrm{L}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$, третий столбец: $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 100.0, \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \lambda^{\mathrm{L}}_{\mathrm{prior}} = 0.01, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$. Первая строка: $\lambda_{\mathrm{temp}} = 0.2$, вторая строка: $\lambda_{\mathrm{temp}} = 1.0$, третья строка: $\lambda_{\mathrm{temp}} = 10.0$.




(а) (б) (в)

Рис. 2.8. Относительная плотность параметров итоговых моделей при $\lambda_{\mathrm{temp}} = 0.2$. Первый столбец: $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 0.01, \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \lambda^{\mathrm{L}}_{\mathrm{prior}} = 100, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$, второй столбец: $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = \lambda^{\mathrm{L}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$, третий столбец: $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 100.0, \lambda^{\mathrm{Q}}_{\mathrm{prior}} = 1, \lambda^{\mathrm{L}}_{\mathrm{prior}} = 0.01, \boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$. Первые 12 параметров соответствуют модели $\mathbf{f}_2$, параметры 13-15 соответствуют модели $\mathbf{f}_1$, параметр 16 соответствует константной модели $\mathbf{f}_0$.

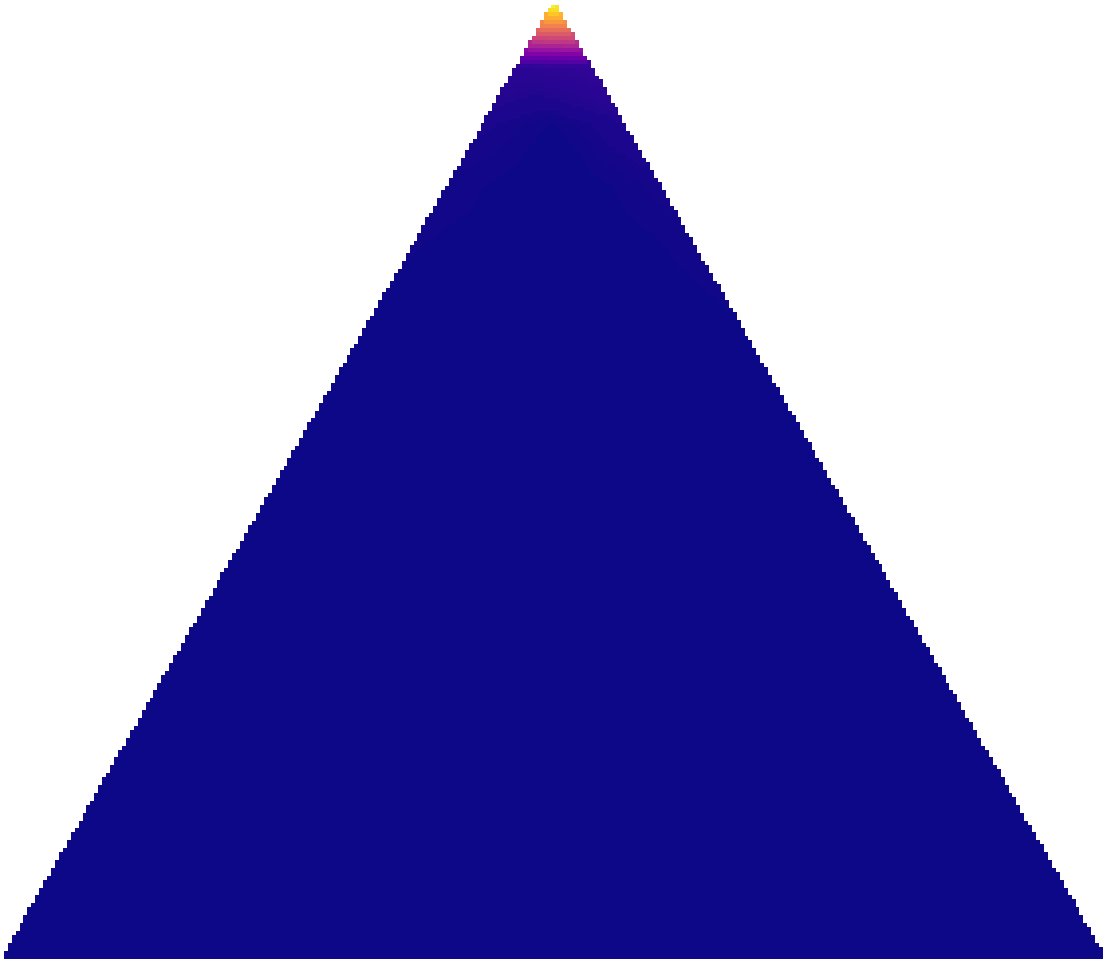


Рис. 2.9. Оптимизация с метапраметром $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = 1$.

## Заключение

Основные результаты диссертационной работы заключаются в следующем.

В главе 1 введены основные понятия, поставлены задачи выбора модели глубокого обучения и проанализированы методы оптимизации параметров модели, методы оптимизации гиперпараметров, методы представления моделей глубокого обучения в графовом виде, методы оптимизации структурных параметров и метапараметров модели. Последние включают в себя как эвристические методы, так и методы, основанные на байесовском выводе и вероятностных предположениях о распределении параметров, гиперпараметров и метапараметров модели.

В главе 2 были предложены критерии оптимальной и субоптимальной сложности моделей глубокого обучения. Предложен алгоритм выбора субоптимальной модели, основанный на получении вариационной нижней оценки правдоподобия модели. Был предложен метод получения оценки, основанный на стохастическом градиентном спуске, позволяющий проводить выбор модели и оптимизацию модели единообразно. Исследованы свойства стохастического градиентного спуска, а также оценок правдоподобия, полученных с его использованием. Работа представленного алгоритма проиллюстрирована рядом выборок. Вычислительный эксперимент продемонстрировал значимое влияние априорного распределения на апостериорное распределение параметров модели.

В главе 3 были проанализированы градиентные методы оптимизации гиперпараметров. Предложено обобщение существующих методов на функции потерь и валидации общего вида. Было проведено сравнение двух критериев выбора модели: на основе кросс-валидации и на основе вариационной оценки правдоподобия модели. Экспреименты показали, что градиентные методы оптимизации гиперпараметров являются эффективными в случае, когда число гиперпараметров велико. Также эксперименты показали, что те модели, гиперпараметры и параметры которых были оптимизированы с использованием вариационной оценки правдоподобия модели, имеют меньшую точность классификации, чем те модели, чьи гиперпараметры и параметры были оптимизированы с использованием метода кросс-валидации. В то же время, первые модели оказались более робастными при доабвлении шума в выборку. Модели, чья оптимизация проводилась с использованием вариационной оценки правдоподобия, оказались значительно лучшими на синтетической выборке, когда число объетов в обучающей выборке мало по сравнению с числом параметров. Поэтому вариационная оценка правдоподобия более предпочтительна, когда вероятность переобучения моделей велика или когда проведение кросс-валидации вычислительно затратно.

В главе 4 был предложен обобщенный метод выбора структуры модели субоптимальной сложности. Формализовано понятие структурной и параметрической сложности для вероятностных моделей. Сформулированы требования к вариационным распределениям, введенным на структуре модели. Показано, что

предложенный метод выбора структуры модели обощает такие методы выбора модели как оптимизация модели градиентным методом, оптимизация метода градиентным методом с регуляризацией, оптимизация модели с ранней остановкой для предотрващения переобучения, методы наращивания и прореживания модели, а также полный перебор.

В главе 5 проведен анализ свойств предложенных методов. Описан реализованный программный комплекс, позволяющий осуществлять выбор моделей глубокого обучения. Проведено сравнение предложенных алгоритмов с известными решениями. Предложенные алгоритмы показали более высокие результаты.

## Список основных обозначений

$\mathbf{x}_i \in \mathbf{X}$ — вектор признакового описания $i$-го объекта
$y_i \in \mathbf{y}$ — метка $i$-го объекта
$\mathfrak{D}$ — выборка
$\mathbf{X} \subset \mathbb{X}$ — матрица, содержащая признаковое описание объектов выборки
$\mathbf{y} \subset \mathbb{Y}$ — вектор меток объектов выборки
$m$ — количество объектов в выборке
$n$ — количество признаков в признаковом описании объекта
$\mathbb{X} = \mathbb{R}^m$ — признаковое пространство объектов
$\mathbb{Y}$ — множество меток объектов
$R$ — множество классов в задаче классификации
$r$ — число оптимизаций модели
$(V, E)$ — граф со множеством вершин $V$ и множеством ребер $E$
$\mathbf{g}^{j,k}$ — вектор базовых функций для ребра $(j, k)$
$K^{j,k}$ — мощность вектора базовых функций для ребра $(j, k)$
$\mathbf{agg}_v$ — функция аггрегации для вершины $v$
$\boldsymbol{\gamma}^{j,k}$ — структурный параметр для ребра $(j, k)$
$\Delta^K$ — симплекс на $K$ вершинах
$\bar{\Delta}^K$ — множество вершин симплекса на $K$ вершинах
$\mathfrak{F}$ — параметрическое семейство моделей
$U$ — область определения оптимизационной задачи
$\mathbf{w} \in \mathbb{W}$ — параметры модели
$\mathbb{W}$ — пространство параметров модели
$U_{\mathbf{w}} \subset \mathbb{W}$ — область определения параметров модели
$\boldsymbol{\Gamma} \in \mathbb{\Gamma}$ — структура модели
$\mathbb{\Gamma}$ — множество значений структуры модели
$U_{\boldsymbol{\Gamma}} \subset \mathbb{\Gamma}$ — область определения параметров модели
$\mathbf{h} \in \mathbb{H}$ — гиперпараметры модели
$\mathbb{H}$ — пространство гиперпараметров модели
$U_{\mathbf{h}} \subset \mathbb{H}$ — область определения гиперпараметров
$\boldsymbol{\theta} \in \Theta$ — параметры вариационного распределения
$\Theta$ — пространство параметров вариационного распределения
$U_{\boldsymbol{\theta}} \subset \Theta$ — область определения вариационных параметров модели
$\boldsymbol{\theta}_{\mathbf{w}} \in \Theta_{\mathbf{w}}$ — параметры вариационного распределения, аппроксимирующего апостериорное распределение параметров модели
$\Theta_{\mathbf{w}}$ — пространство параметров вариационного распределения, аппроксимирующего апостериорное распределение параметров модели
$U_{\boldsymbol{\theta}_{\mathbf{w}}} \subset \Theta_{\mathbf{w}}$ — область определения параметров вариационного распределения, аппроксимирующего апостериорное распределение параметров модели
$\boldsymbol{\theta}_{\boldsymbol{\Gamma}} \in \Theta_{\boldsymbol{\Gamma}}$ — параметры вариационного распределения, аппроксимирующего апостериорное распределение структуры модели
$\Theta_{\boldsymbol{\Gamma}}$ —пространство параметров вариационного распределения, аппроксимирую-

щего апостериорное распределение структуры модели
$U_{\boldsymbol{\theta}_\boldsymbol{\Gamma}} \subset \Theta_{\boldsymbol{\Gamma}}$ — область определения параметров вариационного распределения, аппроксимирующего апостериорное распределение структуры модели
$\boldsymbol{\lambda} \in \mathbb{\Lambda}$ — вектор метапараметров
$\mathbb{\Lambda}$ — пространство метапараметров
$U_{\boldsymbol{\lambda}} \subset \mathbb{\Lambda}$ — область определения метапараметров
$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma})$ — правдоподобие выборки
$p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ — априорное распределение параметров и структуры модели
$p(\mathbf{h}|\boldsymbol{\lambda})$ — распределение гиперпараметров модели
$p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ — априорное распределение структуры модели
$p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ — априорное распределение параметров модели
$p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ — апостериорное распределение параметров и структуры модели
$p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ — апостериорное распределение структуры модели
$p(\boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ — апостериорное распределение структуры модели
$p(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\lambda})$ — апостериорное распределение гиперпараметров
$p(y, \mathbf{w}, \boldsymbol{\Gamma}|\mathbf{x}, \mathbf{h})$ — вероятностная модель глубокого обучения
$p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ — обоснованность модели
$q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta})$ — вариационное распределение параметров и структуры модели
$q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ — вариационное распределение структуры модели
$q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ — вариационное распределение параметров модели
$L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ — функция потерь
$Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ — валидационная функция
$T(\boldsymbol{\theta}|L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}))$ — оператор оптимизации
$\mathfrak{Q}$ — семейство вариационные распределений
$\mathsf{S}$ — энтропия распределения
$M$ — множество моделей без общей параметризации
$D_{\mathrm{KL}}\big(p_1||p_2\big)$ — дивергенция Кульбака-Лейблера между распределениями $p_1$ и $p_2$
$\mathbf{A}^{-1}$ — матрица ковариаций параметров модели
$\mathbf{s}$ — конкатенация параметров концентрации на структуре модели

## Список иллюстраций

# Список таблиц

## Список использованных источников


1. *Grünwald Peter.* A Tutorial Introduction to the Minimum Description Length Principle // Advances in Minimum Description Length: Theory and Applications. — MIT Press, 2005.
2. *Bishop Christopher M.* Pattern Recognition and Machine Learning (Information Science and Statistics). — Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
3. *Salakhutdinov Ruslan, Hinton Geoffrey E.* Learning a Nonlinear Embedding by Preserving Class Neighbourhood Structure // Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07) / Ed. by Marina Meila, Xiaotong Shen. — Vol. 2. — Journal of Machine Learning Research - Proceedings Track, 2007. — Pp. 412–419. http://jmlr.csail.mit.edu/proceedings/papers/v2/salakhutdinov07a/salakhutdinov07a.pdf.
4. On the importance of initialization and momentum in deep learning / Ilya Sutskever, James Martens, George E. Dahl, Geoffrey E. Hinton // Proceedings of the 30th International Conference on Machine Learning (ICML-13) / Ed. by Sanjoy Dasgupta, David Mcallester. — Vol. 28. — JMLR Workshop and Conference Proceedings, 2013. — Май. — Pp. 1139–1147. http://jmlr.org/proceedings/papers/v28/sutskever13.pdf.
5. Approximation and learning by greedy algorithms / Andrew R. Barron, Albert Cohen, Wolfgang Dahmen, Ronald A. DeVore // *Ann. Statist.* — 2008. — 02. — Vol. 36, no. 1. — Pp. 64–94. http://dx.doi.org/10.1214/009053607000000631.
6. *Tzikas Dimitris, Likas Aristidis.* An Incremental Bayesian Approach for Training Multilayer Perceptrons // Artificial Neural Networks – ICANN 2010: 20th International Conference, Thessaloniki, Greece, September 15-18, 2010, Proceedings, Part I / Ed. by Konstantinos Diamantaras, Wlodek Duch, Lazaros S. Iliadis. — Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. — Pp. 87–96. http://dx.doi.org/10.1007/978-3-642-15819-3_12.
7. *Tipping Michael E.* Sparse Bayesian Learning and the Relevance Vector Machine // *J. Mach. Learn. Res.* — 2001. — Сентябрь. — Vol. 1. — Pp. 211–244. http://dx.doi.org/10.1162/15324430152748236.
8. *Cun Yann Le, Denker John S., Solla Sara A.* Optimal Brain Damage // Advances in Neural Information Processing Systems. — Morgan Kaufmann, 1990. — Pp. 598–605.
9. *Попова М. С., Стрижов В. В.* Выбор оптимальной модели классификации физической активности по измерениям акселерометра // *Информатика и ее применения.* — 2015. — Т. 9(1). — С. 79–89. http://strijov.com/papers/Popova2014OptimalModelSelection.pdf.
10. Learning both Weights and Connections for Efficient Neural Network / Song Han, Jeff Pool, John Tran, William Dally // Advances in Neural Information Processing Systems 28 / Ed. by C. Cortes,

N. D. Lawrence, D. D. Lee et al. — Curran Associates, Inc., 2015. — Pp. 1135–1143. http://papers.nips.cc/paper/5784-learning-both-weights-and-connections-for-efficient-neural-network.pdf.

11. Greedy Layer-Wise Training of Deep Networks / Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle // Advances in Neural Information Processing Systems 19 / Ed. by B. Schölkopf, J. C. Platt, T. Hoffman. — MIT Press, 2007. — Pp. 153–160. http://papers.nips.cc/paper/3048-greedy-layer-wise-training-of-deep-networks.pdf.
12. *Hinton Geoffrey E., Osindero Simon, Teh Yee-Whye.* A Fast Learning Algorithm for Deep Belief Nets // *Neural Comput.* — 2006. — Июль. — Vol. 18, no. 7. — Pp. 1527–1554. http://dx.doi.org/10.1162/neco.2006.18.7.1527.
13. Semi-supervised Learning with Deep Generative Models / Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, Max Welling // Advances in Neural Information Processing Systems 27 / Ed. by Z. Ghahramani, M. Welling, C. Cortes et al. — Curran Associates, Inc., 2014. — Pp. 3581–3589. http://papers.nips.cc/paper/5352-semi-supervised-learning-with-deep-generative-models.pdf.
14. *Li Yi, Shapiro L. O., Bilmes J. A.* A generative/discriminative learning algorithm for image classification // Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. — Vol. 2. — 2005. — Oct. — Pp. 1605–1612 Vol. 2.
15. *J. Lasserre.* Hybrid of generative and discriminative methods for machine learning: Ph.D. thesis / University of Cambridge. — 2008.
16. AdaNet: Adaptive Structural Learning of Artificial Neural Networks / Corinna Cortes, Xavier Gonzalvo, Vitaly Kuznetsov et al. // International Conference on Machine Learning. — 2017. — Pp. 874–883.
17. *Zoph Barret, Le Quoc V.* Neural architecture search with reinforcement learning // *arXiv preprint arXiv:1611.01578.* — 2016.
18. Accelerating neural architecture search using performance prediction / Bowen Baker, Otkrist Gupta, Ramesh Raskar, Nikhil Naik // *CoRR, abs/1705.10823.* — 2017.
19. Efficient Architecture Search by Network Transformation / Han Cai, Tianyao Chen, Weinan Zhang et al. — 2018.
20. Learning transferable architectures for scalable image recognition / Barret Zoph, Vijay Vasudevan, Jonathon Shlens, Quoc V Le // *arXiv preprint arXiv:1707.07012.* — 2017.
21. *Liu Hanxiao, Simonyan Karen, Yang Yiming.* Darts: Differentiable architecture search // *arXiv preprint arXiv:1806.09055.* — 2018.
22. *Cho Kyunghyun.* Foundations and Advances in Deep Learning: G5 Artikkeliväitöskirja. — Aalto University; Aalto-yliopisto, 2014. — P. 277. http://urn.fi/URN:ISBN:978-952-60-5575-6.

23. *Alain Guillaume, Bengio Yoshua.* What regularized auto-encoders learn from the data-generating distribution // *Journal of Machine Learning Research.* — 2014. — Vol. 15, no. 1. — Pp. 3563–3593. http://dl.acm.org/citation.cfm?id=2750359.
24. *Kamyshanska Hanna, Memisevic Roland.* On autoencoder scoring // Proceedings of the 30th International Conference on Machine Learning (ICML-13) / Ed. by Sanjoy Dasgupta, David Mcallester. — Vol. 28. — JMLR Workshop and Conference Proceedings, 2013. — Май. — Pp. 720–728. http://jmlr.org/proceedings/papers/v28/kamyshanska13.pdf.
25. *D. Kingma M. Welling.* Auto-Encoding Variational Bayes // Proceedings of the International Conference on Learning Representations (ICLR). — 2014.
26. How to Train Deep Variational Autoencoders and Probabilistic Ladder Networks. / Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe et al. // *CoRR.* — 2016. — Vol. abs/1602.02282. http://dblp.uni-trier.de/db/journals/corr/corr1602.html#SonderbyRMSW16.
27. Semi-Supervised Learning with Ladder Network. / Antti Rasmus, Harri Valpola, Mikko Honkala et al. // *CoRR.* — 2015. — Vol. abs/1507.02672. http://dblp.uni-trier.de/db/journals/corr/corr1507.html#RasmusVHBR15.
28. *MacKay David J. C.* Information Theory, Inference & Learning Algorithms. — New York, NY, USA: Cambridge University Press, 2002.
29. *Токмакова А. А., Стрижов В. В.* Оценивание гиперпараметров линейных и регрессионных моделей при отборе шумовых и коррелирующих признаков // *Информатика и её применения.* — 2012. — Т. 6(4). — С. 66–75. http://strijov.com/papers/Tokmakova2011HyperParJournal_Preprint.pdf.
30. *Зайцев А. А., Стрижов В. В., Токмакова А. А.* Оценка гиперпараметров регрессионных моделей методом максимального правдоподобия // *Информационные технологии.* — 2013. — Vol. 2. — Pp. 11–15. http://strijov.com/papers/ZaytsevStrijovTokmakova2012Likelihood_Preprint.pdf.
31. *Strijov V., Weber Gerhard-Wilhelm.* NONLINEAR REGRESSION MODEL GENERATION USING HYPERPARAMETERS OPTIMIZATION: Preprint 2009-21. — Middle East Technical University, 06800 Ankara, Turkey: Institute of Applied Mathematics, 2009. — Октябрь. — Preprint No. 149.
32. *Стрижов В. В.* Порождение и выбор моделей в задачах регрессии и классификации: Ph.D. thesis / Вычислительный центр РАН. — 2014. http://strijov.com/papers/Strijov2015ModelSelectionRu.pdf.
33. *Перекрестенко Д.О.* Анализ структурной и статистической сложности суперпозиции нейронных сетей. — 2014. http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Perekrestenko2014Comple
34. *Vladislavleva E.* Other publications TiSEM: : Tilburg University, School of Economics and Management, 2008. http://EconPapers.repec.org/RePEc:tiu:tiutis:65a72d10-6b09-443f-8cb9-88f3bb3bc31b.

35. Predicting Parameters in Deep Learning / Misha Denil, Babak Shakibi, Laurent Dinh et al. // Advances in Neural Information Processing Systems 26 / Ed. by C.j.c. Burges, L. Bottou, M. Welling et al. — 2013. — Pp. 2148–2156. http://media.nips.cc/nipsbooks/nipspapers/paper_files/nips26/1053.pdf.
36. *Xu Huan, Mannor Shie.* Robustness and generalization // *Machine Learning.* — 2012. — Vol. 86, no. 3. — Pp. 391–423. http://dx.doi.org/10.1007/s10994-011-5268-1.
37. Intriguing properties of neural networks. / Christian Szegedy, Wojciech Zaremba, Ilya Sutskever et al. // *CoRR.* — 2013. — Vol. abs/1312.6199. http://dblp.uni-trier.de/db/journals/corr/corr1312.html#SzegedyZSBEGF13.
38. Stochastic Variational Inference / Matthew D. Hoffman, David M. Blei, Chong Wang, John Paisley // *J. Mach. Learn. Res.* — 2013. — Май. — Vol. 14, no. 1. — Pp. 1303–1347. http://dl.acm.org/citation.cfm?id=2502581.2502622.
39. *Graves Alex.* Practical Variational Inference for Neural Networks // Advances in Neural Information Processing Systems 24 / Ed. by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett et al. — Curran Associates, Inc., 2011. — Pp. 2348–2356. http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf.
40. *Salimans Tim, Kingma Diederik P., Welling Max.* Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. // ICML / Ed. by Francis R. Bach, David M. Blei. — Vol. 37 of *JMLR Proceedings.* — JMLR.org, 2015. — Pp. 1218–1226. http://dblp.uni-trier.de/db/conf/icml/icml2015.html#SalimansKW15.
41. *Maclaurin Dougal, Duvenaud David K., Adams Ryan P.* Early Stopping is Nonparametric Variational Inference // *CoRR.* — 2015. — Vol. abs/1504.01344. http://arxiv.org/abs/1504.01344.
42. *Mandt Stephan, Hoffman Matthew D, Blei David M.* Continuous-Time Limit of Stochastic Gradient Descent Revisited.
43. *Welling Max, Teh Yee Whye.* Bayesian Learning via Stochastic Gradient Langevin Dynamics // Proceedings of the 28th International Conference on Machine Learning (ICML-11) / Ed. by Lise Getoor, Tobias Scheffer. — ICML '11. — New York, NY, USA: ACM, 2011. — June. — Pp. 681–688.
44. *Arlot Sylvain, Celisse Alain.* A survey of cross-validation procedures for model selection // *Statist. Surv.* — 2010. — Vol. 4. — Pp. 40–79. http://dx.doi.org/10.1214/09-SS054.
45. Fast and Accurate Support Vector Machines on Large Scale Systems / Abhinav Vishnu, Jeyanthi Narasimhan, Lawrence Holder et al. // 2015 IEEE International Conference on Cluster Computing, CLUSTER 2015, Chicago, IL, USA, September 8-11, 2015. — 2015. — Pp. 110–119. http://dx.doi.org/10.1109/CLUSTER.2015.26.

46. Cross-validation pitfalls when selecting and assessing regression and classification models / Damjan Krstajic, Ljubomir J. Buturovic, David E. Leahy, Simon Thomas // *Journal of Cheminformatics*. — 2014. — Vol. 6, no. 1. — Pp. 1–15. http://dx.doi.org/10.1186/1758-2946-6-10.
47. *Hornung Roman, Bernau Christoph, Truntzer Caroline et al.* Full versus incomplete cross-validation: measuring the impact of imperfect separation between training and test sets in prediction error estimation. — 2014. http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-20682-6.
48. *Bengio Yoshua, Grandvalet Yves*. No Unbiased Estimator of the Variance of K-Fold Cross-Validation // *J. Mach. Learn. Res.* — 2004. — Декабрь. — Vol. 5. — Pp. 1089–1105. http://dl.acm.org/citation.cfm?id=1005332.1044695.
49. *Maclaurin Dougal, Duvenaud David, Adams Ryan.* Gradient-based Hyperparameter Optimization through Reversible Learning // Proceedings of the 32nd International Conference on Machine Learning (ICML-15) / Ed. by David Blei, Francis Bach. — JMLR Workshop and Conference Proceedings, 2015. — Pp. 2113–2122. http://jmlr.org/proceedings/papers/v37/maclaurin15.pdf.
50. *Domke Justin.* Generic Methods for Optimization-Based Modeling. // AISTATS / Ed. by Neil D. Lawrence, Mark A. Girolami. — Vol. 22 of *JMLR Proceedings*. — JMLR.org, 2012. — Pp. 318–326. http://dblp.uni-trier.de/db/journals/jmlr/jmlrp22.html#Domke12.
51. *Pedregosa Fabian.* Hyperparameter optimization with approximate gradient // Proceedings of the 33nd International Conference on Machine Learning (ICML). — 2016. http://jmlr.org/proceedings/papers/v48/pedregosa16.html.
52. Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters / Jelena Luketina, Tapani Raiko, Mathias Berglund, Klaus Greff // Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016 / Ed. by Maria-Florina Balcan, Kilian Q. Weinberger. — Vol. 48 of *JMLR Workshop and Conference Proceedings*. — JMLR.org, 2016. — Pp. 2952–2960.
53. *Karaletsos Theofanis, Rätsch Gunnar.* Automatic Relevance Determination For Deep Generative Models // *arXiv preprint arXiv:1505.07765*. — 2015.
54. A monolingual approach to detection of text reuse in Russian-English collection / Oleg Bakhteev, Rita Kuznetsova, Alexey Romanov, Anton Khritankov // Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), 2015 / IEEE. — 2015. — Pp. 3–10.
55. *Бахтеев Олег Юрьевич.* Выбор модели глубокого обучения субоптимальной сложности с использованием вариационной оценки правдоподобия // Интеллектуализация обработки информации ИОИ-2016. — 2016. — Pp. 16–17.

56. Machine-Translated Text Detection in a Collection of Russian Scientific Papers / Alexey Romanov, Rita Kuznetsova, Oleg Bakhteev, Anton Khritankov // *Dialogue*. — 2016. — P. 2.
57. *Бахтеев Олег Юрьевич*. Градиентные методы оптимизации гиперпараметров моделей глубокого обучения // Всероссийская конференция ММРО-18. — 2017. — Pp. 10–11.
58. *Бахтеев Олег Юрьевич, Кузнецова Маргарита Валерьевна*. Детектирование переводных заимствований в текстах научных статей из журналов, входящих в РИНЦ // Всероссийская конференция ММРО-18. — 2017. — Pp. 128–129.
59. *Бахтеев Олег Юрьевич*. Выбо модели глубокого обучения субоптимальной сложности с использованием вариационной оценки правдоподобия // Интеллектуализация обработки информации ИОИ-2018. — 2016. — Pp. 16–17.
60. *Бахтеев Олег Юрьевич, Стрижов Вадим Викторович*. Выбор моделей глубокого обучения субоптимальной сложности // *Автоматика и телемеханика*. — 2018. — no. 8. — Pp. 129–147.
61. *Bakhteev OY, Strijov VV*. Comprehensive analysis of gradient-based hyperparameter optimization algorithms // *Annals of Operations Research*. — 2019. — Pp. 1–15.
62. *Бахтеев ОЮ*. Восстановление панельной матрицы и ранжирующей модели по метризованной выборке в разнородных шкалах // *Машинное обучение и анализ данных*. — 2006. — Vol. 72, no. 7. — P. 1958.
63. *Бахтеев ОЮ*. Восстановление пропущенных значений в разнородных шкалах с большим числом пропусков // *Машинное обучение и анализ данных*. — 2015. — Vol. 1, no. 11. — Pp. 1484–1499.
64. Learning deep generative models of graphs / Yujia Li, Oriol Vinyals, Chris Dyer et al. // *arXiv preprint arXiv:1803.03324*. — 2018.
65. *Li Jundong, Liu Huan*. Challenges of feature selection for big data analytics // *IEEE Intelligent Systems*. — 2017. — Vol. 32, no. 2. — Pp. 9–15.
66. *Hassibi Babak, Stork David G, Wolff Gregory J*. Optimal brain surgeon and general network pruning // Neural Networks, 1993., IEEE International Conference on / IEEE. — 1993. — Pp. 293–299.
67. Incremental network quantization: Towards lossless cnns with low-precision weights / Aojun Zhou, Anbang Yao, Yiwen Guo et al. // *arXiv preprint arXiv:1702.03044*. — 2017.
68. *Han Song, Mao Huizi, Dally William J*. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding // *arXiv preprint arXiv:1510.00149*. — 2015.
69. Dropout: A simple way to prevent neural networks from overfitting / Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky et al. // *The Journal of Machine Learning Research*. — 2014. — Vol. 15, no. 1. — Pp. 1929–1958.

70. *Louizos Christos, Ullrich Karen, Welling Max.* Bayesian compression for deep learning // Advances in Neural Information Processing Systems. — 2017. — Pp. 3290–3300.
71. *Bergstra James, Bengio Yoshua.* Random search for hyper-parameter optimization // *Journal of Machine Learning Research.* — 2012. — Vol. 13, no. Feb. — Pp. 281–305.
72. Algorithms for hyper-parameter optimization / James S Bergstra, Rémi Bardenet, Yoshua Bengio, Balázs Kégl // Advances in Neural Information Processing Systems. — 2011. — Pp. 2546–2554.
73. *Bengio Yoshua.* Gradient-based optimization of hyperparameters // *Neural computation.* — 2000. — Vol. 12, no. 8. — Pp. 1889–1900.
74. DrMAD: Distilling Reverse-Mode Automatic Differentiation for Optimizing Hyperparameters of Deep Neural Networks / Jie Fu, Hongyin Luo, Jiashi Feng et al. // *arXiv preprint arXiv:1601.00917.* — 2016.
75. *Pedregosa Fabian.* Hyperparameter optimization with approximate gradient // Proceedings of the 33rd International Conference on Machine Learning. — 2016.
76. *Snoek Jasper, Larochelle Hugo, Adams Ryan P.* Practical bayesian optimization of machine learning algorithms // Advances in neural information processing systems. — 2012. — Pp. 2951–2959.
77. Bayesian Optimization in High Dimensions via Random Embeddings. / Ziyu Wang, Masrour Zoghi, Frank Hutter et al. // IJCAI. — 2013. — Pp. 1778–1784.
78. Bayesian Optimization with Tree-structured Dependencies / Rodolphe Jenatton, Cedric Archambeau, Javier González, Matthias Seeger // International Conference on Machine Learning. — 2017. — Pp. 1655–1664.
79. Hyperparameter optimization of deep neural networks using non-probabilistic RBF surrogate model / Ilija Ilievski, Taimoor Akhtar, Jiashi Feng, Christine Annette Shoemaker // *arXiv preprint arXiv:1607.08316.* — 2016.
80. Scalable Bayesian Optimization Using Deep Neural Networks / Jasper Snoek, Oren Rippel, Kevin Swersky et al. // Proceedings of the 32nd International Conference on Machine Learning / Ed. by Francis Bach, David Blei. — Vol. 37 of *Proceedings of Machine Learning Research.* — Lille, France: PMLR, 2015. — 07–09 Jul. — Pp. 2171–2180. http://proceedings.mlr.press/v37/snoek15.html.
81. Structure Optimization for Deep Multimodal Fusion Networks using Graph-Induced Kernels / Dhanesh Ramachandram, Michal Lisicki, Timothy J Shields et al. // *arXiv preprint arXiv:1707.00750.* — 2017.
82. Raiders of the lost architecture: Kernels for Bayesian optimization in conditional parameter spaces / Kevin Swersky, David Duvenaud, Jasper Snoek et al. // *arXiv preprint arXiv:1409.4011.* — 2014.
83. *Воронцов Константин Вячеславович.* Локальные базисы в алгебраическом подходе к проблеме распознавания: Ph.D. thesis. — Graz, 1999.

84. *Abadi Martín, Agarwal Ashish, Barham Paul et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. — 2015. — Software available from tensorflow.org. http://tensorflow.org/.
85. *Theano Development Team.* Theano: A Python framework for fast computation of mathematical expressions // *arXiv e-prints.* — 2016. — may. — Vol. abs/1605.02688. http://arxiv.org/abs/1605.02688.
86. Automatic differentiation in PyTorch / Adam Paszke, Sam Gross, Soumith Chintala et al. — 2017.
87. *Eibe Frank, Hall MA, Witten IH.* The WEKA Workbench. Online Appendix for"Data Mining: Practical Machine Learning Tools and Techniques // *Morgan Kaufmann.* — 2016.
88. *Hofmann Markus, Klinkenberg Ralf.* RapidMiner: Data mining use cases and business analytics applications. — CRC Press, 2013.
89. Scikit-learn: Machine learning in Python / Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort et al. // *Journal of machine learning research.* — 2011. — Vol. 12, no. Oct. — Pp. 2825–2830.
90. Relational inductive biases, deep learning, and graph networks / Peter W Battaglia, Jessica B Hamrick, Victor Bapst et al. // *arXiv preprint arXiv:1806.01261.* — 2018.
91. *Negrinho Renato, Gordon Geoff.* Deeparchitect: Automatically designing and training deep architectures // *arXiv preprint arXiv:1704.08792.* — 2017.
92. Learning Bayesian network structure using LP relaxations / Tommi Jaakkola, David Sontag, Amir Globerson, Marina Meila // Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. — 2010. — Pp. 358–365.
93. *Alvarez-Melis David, Jaakkola Tommi S.* Tree-structured decoding with doubly-recurrent neural networks. — 2016.
94. *Adams Ryan, Wallach Hanna, Ghahramani Zoubin.* Learning the structure of deep sparse graphical models // Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. — 2010. — Pp. 1–8.
95. *Feng Jiashi, Darrell Trevor.* Learning the structure of deep convolutional networks // Proceedings of the IEEE international conference on computer vision. — 2015. — Pp. 2749–2757.
96. *Shirakawa Shinichi, Iwata Yasushi, Akimoto Youhei.* Dynamic Optimization of Neural Network Structures Using Probabilistic Modeling // *arXiv preprint arXiv:1801.07650.* — 2018.
97. Toward Optimal Run Racing: Application to Deep Learning Calibration / Olivier Bousquet, Sylvain Gelly, Karol Kurach et al. // *arXiv preprint arXiv:1706.03199.* — 2017.
98. Learning deep resnet blocks sequentially using boosting theory / Furong Huang, Jordan Ash, John Langford, Robert Schapire // *arXiv preprint arXiv:1706.04964.* — 2017.

99. Progressive neural architecture search / Chenxi Liu, Barret Zoph, Jonathon Shlens et al. // *arXiv preprint arXiv:1712.00559*. — 2017.
100. *Alain Guillaume, Bengio Yoshua.* Understanding intermediate layers using linear classifier probes // *arXiv preprint arXiv:1610.01644*. — 2016.
101. *Teerapittayanon Surat, McDanel Bradley, Kung HT.* Branchynet: Fast inference via early exiting from deep neural networks // Pattern Recognition (ICPR), 2016 23rd International Conference on / IEEE. — 2016. — Pp. 2464–2469.
102. Incremental Training of Deep Convolutional Neural Networks / R Istrate12, ACI Malossi, C Bekas, D Nikolopoulos.
103. *Chen Tianqi, Goodfellow Ian, Shlens Jonathon.* Net2net: Accelerating learning via knowledge transfer // *arXiv preprint arXiv:1511.05641*. — 2015.
104. Forward thinking: Building and training neural networks one layer at a time / Chris Hettinger, Tanner Christensen, Ben Ehlert et al. // *arXiv preprint arXiv:1706.02480*. — 2017.
105. *Miranda Conrado S, Von Zuben Fernando J.* Reducing the Training Time of Neural Networks by Partitioning // *arXiv preprint arXiv:1511.02954*. — 2015.
106. *Schmidhuber Juergen, Zhao Jieyu, Wiering MA.* Simple principles of metalearning // *Technical report IDSIA*. — 1996. — Vol. 69. — Pp. 1–23.
107. *Schmidhuber Jürgen.* A neural network that embeds its own meta-levels // Neural Networks, 1993., IEEE International Conference on / IEEE. — 1993. — Pp. 407–412.
108. Meta-SGD: Learning to Learn Quickly for Few Shot Learning / Zhenguo Li, Fengwei Zhou, Fei Chen, Hang Li // *arXiv preprint arXiv:1707.09835*. — 2017.
109. *Wang Yu-Xiong, Hebert Martial.* Learning to learn: Model regression networks for easy small sample learning // European Conference on Computer Vision / Springer. — 2016. — Pp. 616–634.
110. Learning to learn by gradient descent by gradient descent / Marcin Andrychowicz, Misha Denil, Sergio Gomez et al. // Advances in Neural Information Processing Systems. — 2016. — Pp. 3981–3989.
111. *Kinga D, Adam J Ba.* A method for stochastic optimization // International Conference on Learning Representations (ICLR). — Vol. 5. — 2015.
112. *Duchi John, Hazan Elad, Singer Yoram.* Adaptive subgradient methods for online learning and stochastic optimization // *Journal of Machine Learning Research*. — 2011. — Vol. 12, no. Jul. — Pp. 2121–2159.
113. *Friesen Abram L, Domingos Pedro.* Deep Learning as a Mixed Convex-Combinatorial Optimization Problem // *arXiv preprint arXiv:1710.11573*. — 2017.
114. *Kristiansen Gus, Gonzalvo Xavi.* EnergyNet: Energy-based Adaptive Structural Learning of Artificial Neural Network Architectures // *arXiv preprint arXiv:1711.03130*. — 2017.

115. Pathnet: Evolution channels gradient descent in super neural networks / Chrisantha Fernando, Dylan Banarse, Charles Blundell et al. // *arXiv preprint arXiv:1701.08734*. — 2017.
116. *Veniat Tom, Denoyer Ludovic.* Learning time-efficient deep architectures with budgeted super networks // *arXiv preprint arXiv:1706.00046*. — 2017.
117. Composing graphical models with neural networks for structured representations and fast inference / Matthew Johnson, David K Duvenaud, Alex Wiltschko et al. // Advances in neural information processing systems. — 2016. — Pp. 2946–2954.
118. *Nalisnick Eric, Smyth Padhraic.* Deep Generative Models with Stick-Breaking Priors // *arXiv preprint arXiv:1605.06197*. — 2016.
119. *Abbasnejad M Ehsan, Dick Anthony, van den Hengel Anton.* Infinite variational autoencoder for semi-supervised learning // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) / IEEE. — 2017. — Pp. 781–790.
120. *Miller A. C., Foti N., Adams R. P.* Variational Boosting: Iteratively Refining Posterior Approximations // *ArXiv e-prints*. — 2016. — nov.
121. *Arnold Ludovic, Ollivier Yann.* Layer-wise learning of deep generative models // *arXiv preprint arXiv:1212.1524*. — 2012.
122. *Maddison Chris J, Mnih Andriy, Teh Yee Whye.* The concrete distribution: A continuous relaxation of discrete random variables // *arXiv preprint arXiv:1611.00712*. — 2016.
123. On some variance reduction properties of the reparameterization trick / Ming Xu, Matias Quiroz, Robert Kohn, Scott A Sisson // *arXiv preprint arXiv:1809.10330*. — 2018.
124. The Reparameterization Trick. http://gregorygundersen.com/blog/2018/04/29/reparamete
125. *Hinton Geoffrey, Van Camp Drew.* Keeping neural networks simple by minimizing the description length of the weights // in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory / Citeseer. — 1993.