ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

На правах рукописи

Струминский Кирилл Алексеевич

ГАРАНТИИ ОБУЧЕНИЯ И ЭФФЕКТИВНЫЙ ВЫВОД В ЗАДАЧАХ СТРУКТУРНОГО ПРЕДСКАЗАНИЯ

РЕЗЮМЕ
диссертации на соискание учёной степени
кандидата компьютерных наук

Москва — 2023

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики».

Научный руководитель: Ветров Дмитрий Петрович, к.ф.-м.н., Национальный исследовательский университет «Высшая школа экономики».

1 Введение

Машинное обучение пытается восстановить и описать эмпирические зависимости в данных. Часто интерес представляют количественные показатели или атрибуция наблюдаемых данных к заранее заданному набору категорий. Например, как зависит цена квартиры от её расположения и параметров? Захочет ли пользователь читать данное письмо? На эти вопросы можно ответить, на основе исторических данных, содержащих детали прошлых сделок или историю взаимодействия пользователя с полученными ранее письмами. Также интерес может представлять и атрибуция, когда атрибуты не известны заранее: можно ли, например, в данных выделить несколько характерных категорий?

В то же время, в приложениях возникают задачи, в которых искомые зависимости выпадают за рамки описанных выше примеров. Если, например, речь идёт о задаче машинного перевода, то каждому тексту исходного языка необходимо сопоставить текст на выходном языке. Предсказываемый перевод в этом случае было бы некорректно представлять в виде числа или элемента множества всех возможных переводов. Напротив, текст было бы удобно представить как последовательность слов, где алгоритм перевода должен предсказать каждое слово, ориентируясь и на исходное предложение, и на соседние слова перевода.

Переменные в данных, представимые в виде совокупности взаимно-зависимые величин, принято называть структурными. Подраздел машинного обучения, связанный с предсказанием структурных величин, принято называть структурным предсказанием. Характерной особенностью структурных переменных является комбинаторный рост числа возможных значений в зависимости от параметров задачи. Опуская нюансы задачи, в примере с машинным переводом при размере словаря w и заранее известной длиной перевода l алгоритм должен выбрать среди w^l возможных переводов. Эта особенность делает актуальными вопросы о гарантиях обучения и эффективном выводе. А именно, сколько примеров достаточно чтобы надежно восстановить искомую зависимость? Как быстро выбирать элемент из возможного множества исходов? Изучению этих вопросов и посвящена данная работа.

1.1 Актуальность темы исследования

Вместе с ростом област применения машинного обучения[56], растёт и спектр решаемых задач, делая более востребованным и задачи структурного предсказания. В частности, развитие глубинного обучения позволило вывести на качественно новый уровень алгоритмы обработки естественного языка и компьютерного зрения. В задачах обучения с учителем, где в качестве целевых переменных выступают структурные переменные, обучение часто сводят к минимизации кросс-энтропийной функции потерь. Для этого вводят распределение структурной переменной. Например, в задачах обработки естественного языка, распределение над текстовыми выходами вводят, как правило, разбивая распределение на соответствующие словам компоненты по цепному правилу (подробнее в [23, Глава 10]). Другое решение, распространенное, например, в задаче семантической сегментации, заключается в том чтобы предположить все элементы структурной переменной независимыми при условии входного изображения (как, например, сделано в работе [54]). Исследования последних лет, как правило, посвящены оптимизации архитектур нейронных сетей для параметризации распределений в описанном подходе, а также масштабированию описанного подхода [8, 30, 72]. Необходимость предсказания структурных переменных подтолк-

нула развитие рекуррентных [28, 62] и сверточных нейронных сетей [22, 43], а также трансформеров [67] для обработки последовательностей, архитектуры UNet для обработки изображений [54].

Недостатком описанного выше подхода к структурному предсказанию и глубинного обучения в целом является ограниченная интерпретируемость найденных зависимостей. В то же время, в приложениях было введено понятие "права на объяснение"[68], согласно которому человек может потребовать объяснение того каким образом система машинного обучения приняла решение относительно него. Таким образом, проблема интерпретации алгоритмов машинного обучения встает особенно остро с развитием систем глубинного обучения. Для решения этой проблемы появились работы по интерпретации конкретных архитектур [69, 31, 52], а также подходы к интерпретации произвольных алгоритмов машинного обучения [37, 51, 11]. В то же время, распространение получила идея о применении скрытых структурных переменных для повышения интерпретируемости алгоритмов машинного обучения [32, 39]. Далее подробнее опишем идею. Глубинные нейронные сети представляют из себя последовательность элементарных вычислительных блоков, в то время как совокупный вычисления этих блоков с трудом поддается интерпретации. С другой стороны, если у каких-то из этих промежуточных блоков по построению будет интерпретируемый (структурный) выход, а сама архитектура сети подобрана с учетом специфики задачи, интерпретация результата вычисления сети может оказаться проще. Так, например, при анализе настроения текста можно выбрать подмножество слов, на основе которых модель затем предскажет настроение текста. На практике, выбранные в такой модели слова помогают проинтерпретировать сделанное предсказание. В то же время, нейронные сети со скрытыми структурными переменными можно рассматривать как развитие моделей со скрытыми переменными, таких как скрытые марковские цепи [12] или вероятностные контекстно свободные грамматики [55] для моделирования языков, за счет добавления более выразительных нейросетевых моделей.

Однако, в случае дискретных скрытых переменных метод обратного распространения ошибки для обучения нейронной сети оказывается неприменим из-за недифференцируемости блока, возвращающего скрытую переменную. Решение этой проблемы, как правило, сводится к эвристическому определению градиентов блока [4] или стохастической релаксации [29, 38, 5, 45]. Часть данного исследования посвящена проблеме обучения со скрытыми перестановками. Другой проблемой, связанной с скрытыми структурными переменными и не теряющей своей актуальности по сей день, является разработка архитектур со скрытыми переменными и выбор целевых функций. Как показывает практика [33, 16], попытки сквозного обучения таких моделей приводят к тому, что предсказательная модель может не опираться на скрытые переменные, выучивая зависимость лишь на основе стандартных нейросетевых компонент. Стандартным решением в этом случае является обучение с частичной разметкой скрытых переменных: для доли обучающих примеров вводится дополнительная функция потерь, поощряющая желаемое предсказание. Альтернативой может быть выбор архитектуры, не позволяющий добиться достаточной точности предсказания без использования скрытой переменной [11].

Наряду с развитием практических подходов и алгоритмов для работы со структурными переменными важной задачей также является получение гарантий качества их работы. Возвращаясь к задаче структурного предсказания отметим, что комбинаторный рост числа возможных предсказаний и неравный вклад ошибочных предсказаний (не все неточные предсказания одинаково плохи) являются двумя факторами,

которые отличают структурное предсказание от подробно изученной задачи классификации [44]. Обобщающей способности в контексте структурного предсказания посвящены работы [17, 36]. Целевая метрика зачастую не совпадает с оптимизируемым при обучении функционалом (суррогатная функция потерь), был получен ряд результатов о их связи для задач структурного предсказания. В работе [14] авторы показали состоятельность широкого класса квадратичных суррогатных функций потерь, а работа [44] получила оценку на расхождение точности предсказания согласно целевой метрике и суррогатной функции потерь. В дальнейшем, эти результаты были обобщены на гладкие выпуклые суррогатные функции потерь [42]. Однако, приведенные выше работы выполнены в предположении состоятельности суррогатной функции потерь, в то время как на практике часто применяются несостоятельные функции потерь: например, многоклассовый метод опорный векторов в форме Краммера-Сингера [19], а также его обобщения на структурные переменные [63, 65]. В рамках исследования несостоятельных функций потерь, данное диссертационное исследование обобщило результаты [44], получив оценки для квадратичных суррогатных функций потерь без дополнительного требования состоятельности.

1.2 Цели и задачи исследования

Как было отмечено выше, практика применения машинного обучения зачастую приводит к необходимости предсказания структурных переменных. Возможные постановки задач при этом могут включать структурные переменные в качестве целевых переменных в случае обучения с учителем, а также в качестве скрытых вспомогательных переменных. Помимо значения целевых метрик предсказания, в силу комбинаторной природы структурных переменных, важным критерием оказывается скорость построения предсказания. Целью данного исследования была разработка методов структурного предсказания, отвечающих возникающим на практике требованиям: разработка методов структурного предсказания с наблюдаемыми и скрытыми структурными переменными, акцент на быстрые алгоритмы предсказания структурной переменной, наличие гарантий обучения у предложенных методов.

В рамках описанной выше цели были поставлены следующие **задачи**:

1. разработка и исследования методов предсказания для таких структурных переменных как перестановки и подмножества заданного размера,
2. исследование состоятельности и построение гарантий обучения для задач обучения с учителем со структурной целевой переменной,
3. разработка и эмпирический анализ методов обучения со скрытыми структурными переменными,
4. разработка эффективных методов вывода структурных переменных
5. применение скрытых структурных переменных для интерпретации данных, а также построения интерпретируемых методов машинного обучения.

Научная новизна. При решении поставленных задач были получены следующие результаты.

1. Разработан и исследован градиентный метод оптимизации по множеству перестановок.

2. Проведен теоретический анализ квадратичных суррогатных функций потерь в задачах обучения с учителем со структурной целевой переменной.

3. Предложен и изучен ряд подходов к восстановлению скрытых структурных переменных с использованием метода максимальной обоснованности, а также квадратичных суррогатных функций потерь.

4. Разработаны эффективные методы вывода структурных переменных для случая перестановок и подмножеств фиксированного размера.

5. Разработаны методы интерпретации данных на основе скрытых структурных переменных.

1.3 Практическая значимость

Разработанный подход к оптимизации по перестановкам применим для восстановления структуры зависимости между переменными в данных, что, в частности, оказывается востребованным при интерпретации моделей машинного обучения. Априорное распределение для параметров сверточной сети предлагает метод быстрой адаптации параметров модели к новому смежному домену данных. Метод оценки параметров многопользовательского канала связи находит применение в современных сотовых сетях. Вероятностная модель для предобработки данных геолого-физической разведки дает удобный способ для детекции аномалий и восстановления пропусков в исторических данных.

1.4 Методология и методы исследования

Теоретический анализ задач структурного предсказания опирается на разделы теории вероятностей, статистической теории обучения, оптимизации. Абстрагированная постановка позволяет получить общий результат для ряда задач структурного предсказания. Для формализации остальных задач были использованы методы вероятностного машинного обучения, а также байесовского подхода к машинному обучению. Предложенные алгоритмы опираются на базовые разделы теории вероятности и стохастической оптимизации. Помимо нескольких формально доказанных утверждений, данная работа по большей части опирается на эмпирический метод исследования. Предложенные алгоритмы были реализованы на языке Python, качество их работы было сопоставлено с аналогами на синтетических и реальных наборах данных.

1.5 Публикации и апробация работы

Публикации повышенного уровня:

1. **Struminsky K.**, Lacoste-Julien S., Osokin A. Quantifying Learning Guarantees for Convex but Inconsistent Surrogates //Advances in Neural Information Processing Systems. – 2018. – С. 669-677. *Вклад автора диссертации:* Нижняя оценка на калибровочную функцию в задачах структурного обучения в общем виде; вычисление коэффициентов оценки калибровочной функции для задачи иерархической классификации; вычисление коэффициентов оценки калибровочной функции для задачи ранжирования.

2. Gadetsky, A., **Struminsky, K.**, Robinson, C., Quadrianto, N., & Vetrov, D. P. (2020). Low-Variance Black-Box Gradient Estimates for the Plackett-Luce Distribution. In AAAI (pp. 10126-10135). *Вклад автора диссертации:* Подход к оптимизации по перестановкам и ациклическим графам на основе вариационной оптимизации для распределений Плакетта-Люса; обобщение алгоритма RELAX на случай распределения Плакетта-Люса.

3. Atanov, A., Ashukha, A., **Struminsky, K.**, Vetrov, D., & Welling, M. (2018, September). The Deep Weight Prior. In International Conference on Learning Representations. *Вклад автора диссертации:* Адаптация вероятностной модели вариационного автокодировщика для задачи оценки априорного распределения на параметры байесовской нейронной сети.

Публикации стандартного уровня:

1. **Struminsky K.** et al. A new approach for sparse Bayesian channel estimation in SCMA uplink systems //2016 8th International Conference on Wireless Communications & Signal Processing (WCSP). – IEEE, 2016. – С. 1-5. *Вклад автора диссертации:* Вероятностная модель для оценки параметров многопользовательского канала связи; улучшенная схема приближенного вывода параметров многопользовательского канала связи и оценки конфигурации канала.

2. **Struminskiy K.** et al. Well Log Data Standardization, Imputation and Anomaly Detection Using Hidden Markov Models //Petroleum Geostatistics 2019. – European Association of Geoscientists & Engineers, 2019. – Т. 2019. – №. 1. – С. 1-5. *Вклад автора диссертации:* Вероятностная модель для предобработки данных геологофизической разведки.

Во всех работах, за исключением работы "The Deep Weight Prior-[1], соискатель является главным автором.

Доклады на конференция и семинарах:

1. Bayesian Deep Learning Workshop, NeurIPS 2019, Ванкувер, Канада, 13 декабря, 2019.

   Тема: Low-variance Gradient Estimates for the Plackett-Luce Distribution (устный доклад, постер).

2. 8th International Conference on Wireless Communications and Signal Processing, Янчжоу, Китай, 13-15 октября, 2016.

   Тема: A new approach for sparse Bayesian channel estimation in SCMA uplink systems (устный доклад).

3. Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), Нью-Йорк, США, 7-12 февраля, 2020.

   Тема: Low-Variance Black-Box Gradient Estimates for the Plackett-Luce Distribution (устный доклад, постер).

4. EAGE Conference on Petroleum Geostatistics, Флоренция, Италия, 2-6 сентября, 2019.

   Тема: Well Log Data Standardization, Imputation and Anomaly Detection Using Hidden Markov Models (устный доклад).

5. Thirty-second Annual Conference on Neural Information Processing Systems (NeurIPS 2018), Монреаль, Канада, 2-8 декабря, 2018.

   Тема: Quantifying Learning Guarantees for Convex but Inconsistent Surrogates (постер).

6. Thirty-fifth Annual Conference on Neural Information Processing Systems (NeurIPS 2021), онлайн, 6-14 декабря, 2021.

   Тема: Leveraging Recursive Gumbel-Max Trick for Approximate Inference in Combinatorial Spaces (постер).

7. Seventh International Conference on Learning Representations (ICLR 2019), Новый Орлеан, США, 6-9 мая, 2019.

   Тема: The Deep Weight Prior (постер).

8. Семинар байесовской группы, Москва, Россия, 26 октября, 2018.

   Тема: Quantifying Learning Guarantees for Convex but Inconsistent Surrogates (устный доклад).

9. Sberbank Data Science Journey, Москва, Россия, 10 ноября, 2018.

   Тема: Quantifying Learning Guarantees for Convex but Inconsistent Surrogates (устный доклад, постер).

10. Machines Can See: Computer Vision and Deep Learning Summit, Москва, Россия, 25 июня, 2019.

    Тема: The Deep Weight Prior (постер).

11. International Conference on Analysis of Images, Social Networks and Texts, AIST 2019, Казань, Россия, 17-19 июля, 2019.

    Тема: A Simple Method to Evaluate Support Size and Non-uniformity of a Decoder-Based Generative Model (устный доклад).

12. Advances in Approximate Bayesian Inference, NIPS 2016 Workshop, Барселона, Испания, 2016.

    Тема: Robust Variational Inference (постер).

2 Содержание работы

2.1 Структурные переменные в машинном обучении

Первым делом введем понятие структурной переменной. В рамках машинного обучения, структурная переменная - собирательный термин для возникающих в задачах величин, объединенных следующими характерными свойствами. Во-первых, большое количество возможных значений: как правило, конечное, но не допускающее быстрого перебора на компьютере. Во-вторых, в задачах эти переменные представлены в виде набора зависимых случайных величин. Второе свойство можно положить общим определением структурное переменной; для большей наглядности мы перейдем к конкретным примерам.

В задачах машинного обучения структурные переменные могут выступать в роли целевой переменной в задачах предсказания (структурного предсказания), а также могут выступать в качестве вспомогательной скрытой переменной в моделях со скрытыми переменными.

Одним из классических примеров задачи структурного предсказания является сегментация в компьютерном зрении [34] В этом случае структурной переменной является маска сегментации изображения. Компоненты макси сегментации зависимы, поскольку близкие точки изображения с высокой вероятностью соответствуют одному и тому уже классу. Другие примеры задач структурного предсказания включают ранжирование [10, 49], экстремальную классификацию [13]. Многие задачи обработки естественного языка также являются задачами структурного предсказания. Модель, выдающая на выходе текст, будь то суммаризация, перевод или ответ на вопрос, должна предсказывать последовательность взаимозависимых случайных величин. В глубинном обучении такие модели определены seq2seq архитектурой [62], а для предсказания используют приближенные алгоритмы поиска среди всех возможных вариантов [50].

До распространения методов глубинного обучения структурные переменные были также востребованы в задачах обработки естественного языка, играя там зачастую роль вспомогательных переменных [58]. Так, например, при переводе могли опираться на дерево зависимостей входного предложения для лучшей передачи смысла. В этом примере структурной переменной является дерево разбора предложения, а для его построения дерева могли использовать отдельную модель, обученную на других данных.

Тем не менее, сегодня подобная практика отошла на второй план. Глубинные нейронные сети допускают сквозное обучение, предобучение на неразмеченных данных [41, 20] и перенос знаний на маленькие выборки [74]. С этой точки зрения сегодня интерес представляют алгоритмы со скрытыми структурными переменным, допускающие сквозное обучение. Они позволяют взять лучшее из двух миров: с одной стороны выразительную способность нейронный сетей, с другой стороны опору на априорные знания за счет структурных переменных для лучшей интерпретируемости и более эффективного использования данных.

Среди примеров моделей со скрытыми структурными переменными можно выделить скрытые марковские цепи [48] с разметкой последовательности в качестве структурной переменной, вероятностные контекстно-свободные грамматики [15] с деревом разбора в качестве структурной переменной, а также модель темпоральной классификации последовательностей [25] со скрытой разметкой входной последовательности. Перечисленные модели опирались на модели специального ограниченного вида, необходимые для быстрой обработки структурной переменной. Альтернативой ограниченному виду моделей служит сквозное обучения на основе стохастического градиентного спуска [45]. Примерами могут служить модели со скрытыми деревьями разбора [16], неявным выбором подмножества признаков [11] и скрытым порядком генерации текста [27].

Переходя от понятия структурной переменной, в следующем разделе мы определим общую задачу структурного предсказания.

2.2 Основы структурного предсказания

Для начала рассмотрим классическую постановку задачи структурного предсказания. А именно, рассмотрим задачу обучения с учителем, где на вход модели посту-

пают объекты $x \in \mathcal{X}$ из множества $\mathcal{X}$, а модель должна предсказать структурную переменную $y \in \mathcal{Y}$, принимающую значения в конечном множестве $\mathcal{Y}$. Данные распределены по закону $\mathcal{D}$, а y является реализацией случайного вектора Y с носителем $\mathcal{Y} \subset \mathbb{R}^m$. В общем случае, метки обучающей выборки могут лежать в отличном от $\mathcal{Y}$ множестве $\hat{\mathcal{Y}}$.

Для описания алгоритма предсказания модели мы зададим функцию $f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$, которая присваивает каждому возможной структуре $y \in \mathcal{Y}$ число, а затем выбирает оптимальную структуру в качестве предсказания

$$\text{pred}(f(x)) := \arg\max_{y \in Y} f_y(x). \tag{1}$$

Отличие структурного предсказания от других задач обучения с учителем заключается в том, что множество возможных исходов $\mathcal{Y}$ велико из-за комбинаторного роста возможных предсказаний. Например, при ранжировании исходом может быть перестановка элементов, а при сегментации последовательность меток классов. Поэтому модель должна предлагать быстрый способ решения задачи 1 при выводе. Помимо этого, функцию f необходимо эффективно хранить в памяти. Как правило, на практике рассматривают малоранговую параметризацию функции $f(x) = Fg(x)$ для фиксированного линейного оператора $F : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$ и функции $g(x) : \mathcal{X} \to \mathbb{R}^d$. Такая параметризация позволяет сократить количество параметров функции, а также упростить задачу вывода 1 благодаря структуре матрицы F. Параметризация также ограничивает множество возможных предсказаний, поскольку вектор предсказаний $f(x)$ обязательно лежит в линейной оболочке столбцов матрицы $\mathcal{F} = \operatorname{span} F$. Пространство $\mathcal{F}$ мы будем называть множеством возможных предсказаний.

Цель обучения - найти функцию f, оптимальную с точки зрения риска (ожидаемого значения функции потерь) $L(\cdot, \cdot) : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}$

$$\mathcal{R}_L(f) := \mathbb{E}_{X,Y} L(\text{pred}(f(X)), Y). \tag{2}$$

Функция потерь оценивает качество предсказаний, но её оптимизация может быть затруднительна. В частности, она не является дифференцируемой функцией от параметров модели $f(X)$, Для оптимизации прибегают к вспомогательной (суррогатной) функции потерь. Введем $\Phi : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$

$$\mathcal{R}_\Phi(f) := \mathbb{E}_{X,Y} \Phi(f(x), y). \tag{3}$$

Можно заметить, что, в отличие от целевой функции 2, функция 3 принимает не предсказания, а напрямую значение $f(x)$, что существенно упрощает оптимизацию. Среди популярных суррогатных функций потерь можно выделить квадратичные функции [14, 7], функции основанные на правдоподобии [34] и вариации целевой функции SVM [53, 65].

При замене целевой функции на суррогатную неизбежно возникает вопрос о связи оптимизации суррогатной функции потерь с решением исходной задачи. Для получения ответа на этот вопрос было введено понятие состоятельности суррогатной функции потерь [2], близко связанное с понятием состоятельности по Фишеру [18, стр. 287]. Интуитивно, суррогатная функция потерь состоятельна, если её оптимизация на генеральной совокупности приводит к оптимальной модели предсказания.

Определим состоятельность через калибровочную функцию, связывающую суррогатную и целевую функции потерь. Для предсказания $f \in \mathcal{F} \subseteq \mathbb{R}^k$ и распределения $q \in \Delta_k$ возможных исходов $Y \sim q$ введем условный риск $l(f, q) := \mathbb{E}_Y L(\text{pred}(f), Y)$

и условный суррогатный риск $\phi(f,q) := \mathbb{E}_Y \Phi(f,Y)$. Избыточным (суррогатным) риском назовем отклонение $\delta l(f,q)$ ($\delta\phi(f,q)$ от наименьшего возможного риска

$$\delta l(f,q) := l(f,q) - \inf_{\hat{f}\in\mathcal{F}} l(\hat{f},q) \tag{4}$$

$$\delta\phi(f,q) := \phi(f,q) - \inf_{\hat{f}\in\mathcal{F}} \phi(\hat{f},q). \tag{5}$$

Используя эти вспомогательные функции, дадим определение калибровочной функции.

Определение 1. Для функции потерь L и суррогатной функции потерь Φ, а также множества возможных предсказаний $\mathcal{F}$ калибровочная функция $H_{\Phi,L,\mathcal{F}}(\varepsilon)$ определена для аргументов $\varepsilon \geq 0$ и равна наименьшему возможной ошибке суррогатного риска при ошибке целевого риска не меньше ε:

$$H_{\Phi,L,\mathcal{F}}(\varepsilon) := \inf_{f\in\mathcal{F}, q\in\Delta_k} \delta\phi(f,q) \tag{6}$$

$$\text{s.t.}\delta l(f,q) \geq \varepsilon \tag{7}$$

Неформально, калибровочная функция оценивает насколько малой может быть ошибка суррогатной функции потерь при данном уровне ошибки целевой функции потерь. Согласованная суррогатная функция потерь не должна принимать низкие значения, если целевая функция потерь принимает высокие значения в результате ошибки предсказания. Следующая теорема позволяет связать суррогатный риск и целевой риск с помощью калибровочной функции.

Теорема 1 (Связь риска с суррогатным риском через калибровочную функцию)**.** *Пусть $H_{\Phi,L,\mathcal{F}}$ калибровочная функция между функцией потерь L и суррогатной функцией потерь Φ, а $\mathcal{F}$ задает множество возможных предсказаний. Пусть $\Phi, \hat{L}, \mathcal{F}$ является выпуклой неубывающей нижней оценкой калибровочной функции $H_{\Phi,L,\mathcal{F}}$. Дополнительное предположим, что Φ непрерывная и ограниченная снизу функция. Тогда для любого $\varepsilon > 0$, такого, что $\hat{H}_{\Phi,L,\mathcal{F}}$ конечна, и любого предсказания $f \in \mathcal{F}$ выполнено*

$$\mathcal{R}_\Phi(f) < \inf_{\hat{f}\in\mathcal{F}} R_\phi(\hat{f}) + \hat{H}_{\Phi,L,\mathcal{F}}(\varepsilon) \Rightarrow \mathcal{R}_L(f) < \inf_{\hat{f}\in\mathcal{F}} \mathcal{R}_L(\hat{f}) + \varepsilon. \tag{8}$$

Следуя теореме, дадим определения η-состоятельности суррогатной функции потерь.

Определение 2 (η-состоятельность суррогатной функции потерь)**.** Суррогатная функция потерь Φ состоятельна вплоть до уровня $\eta \geq 0$ (η-состоятельная) для целевой функции L и множества возможных предсказаний $\mathcal{F}$ тогда, и только тогда, когда для калибровочной функции выполнено $H_{\Phi,L,\mathcal{F}}(\varepsilon) > 0$ для любого $\varepsilon > \eta$ и существует $\hat{\varepsilon} > \eta$, такой, что $H_{\Phi,L,\mathcal{F}}(\hat{\varepsilon})$ конечна.

При $\eta = 0$ данное выше определение совпадает с понятием согласованности, принятым в литературе по машинному обучению [35, 47]. Таким образом, проверка состоятельности сводится к тому чтобы показать положительность калибровочной функции в проколотой окрестности $\varepsilon = 0$. Однако, на практике состоятельность может оказаться недостаточной для построения реалистичных гарантий обучения. Как было показано в работе [44], для многих задач структурного предсказания теорема 1

требует практически недостижимой точности оптимизации суррогатной функции. Важной оказывается не только состоятельность, но и сами значения калибровочной функции.

Определение состоятельности для $\eta > 0$ оказывается важным для результатов, представленных в этой работе. Во-первых, оно позволяет давать более слабые гарантии обучения в теореме 1 для несостоятельных суррогатных функций, то есть не попадающих под определение 0-состоятельности. Во-вторых, для несостоятельных калибровочных функций потерь нам удалось дать более оптимистичные гарантии обучения благодаря более низким значениям калибровочной функции.

2.3 Вероятностные методы структурного подхода

Постановка задачи структурного обучения в разделе 2.2 использовала язык теории вероятностей для введения предположений о данных и сведения обучения к задаче оптимизации. Помимо этого, язык теории вероятностей оказывается полезен при определении недетерминированных моделей предсказания и при описании таких аспектов неопределенности в работе систем машинного обучения, как неопределенность в выборе модели и неопределенность в предсказании конкретной модели. **Вероятностным машинным обучением** принято называть подход к машинному обучению, опирающийся на теорию вероятностей для постановки и решения задач. Далее мы более подробно опишем этот подход, отталкиваясь от распространенных примеров его применения.

При постановке задачи в рамках вероятностного подхода первым делом выделяют набор случайных величин, фигурирующих в задаче, а также их совместное распределение. Трактуя данные как случайные величины, мы описываем искомые закономерности выбирая подходящий класс распределений.

Так, например, при построении модели логистической регрессии, метку класса $y \in \mathcal{Y} = \{-1, 1\}$ представляют как случайную величину Y с распределением Бернулли, зависящем от входного объекта $x \in \mathcal{X} = \mathbb{R}^d$, и вероятностью $\mathbb{P}_Y(y \mid x; \theta) = \frac{1}{1+\exp(y\theta^T x)}$ с параметрами $\theta \in \mathbb{R}^d$. Следуя предположениям о данных, мы полагаем совместную независимость объектов x и меток y. Распределение меток позволяет описать неопределенность в предсказании модели, которая может быть связана с недостатком данных или негибкостью модели. В случае, если неопределенность возникает и при оценке параметров модели, вероятностный подход позволяет рассмотреть в качестве случайной величины Θ также и параметры модели. В отсутствии каких-то знаний о значениях параметра Θ, его распределение можно положить нормальным $\Theta \sim N(\mid 0, \operatorname{diag}\sigma), \sigma \in \mathbb{R}^d$ с диагональной матрицей ковариаций с параметрами $\sigma \in \mathbb{R}^d$. Предполагая $\Theta \perp\!\!\!\perp Y$, мы получим совместное распределение параметров Θ и меток Y. Интерпретация параметров модели в качестве случайных величин лежит в основе байесовского подхода, позволяя оценивать неопределенность при выборе параметров модели с помощью апостериорного распределения $p_\Theta(\theta \mid (x^i, y^i)_{i=1}^n; \sigma)$ для набора данных из n объектов.

Заметим, что приведенный выше пример не делал предположений о распределении входных объектов x, поскольку в них нет необходимости при рассмотрении задачи классификации. Подобные модели называются дискриминативными. В то же время вероятностный подход позволяет противопоставить дискриминативным моделям порождающие, или генеративные, которые также моделируют распределение входных объектов. Классическим примеров порождающей модели может служить наивный байесовский классификатор. В его основе лежит совместное распределение

$p(x, y \mid \theta) := p(x \mid y; \theta)p(y; \theta)$, а при разделении объектов он опирается на условное распределение $p(y \mid x, \theta)$.

Помимо этого, вероятностный подход позволяет вводить дополнительные случайные величины, позволяя упростить описание искомых зависимостей. Так, например, модель латентного размещения Дирихле[6] для текстов группирует объекты $x \in \mathcal{X}$ с соответствии с темами: для корпуса текстов модель определяет набор из $\tau \in \mathbb{N}$ тем, а затем представляет каждый отдельный текст $p(x \mid t)$ на основе вектора тем $t \in \Delta^T$, которые в тексте отражены. Вспомогательной случайной величиной в этом случае оказывается набор тем текста t с априорным распределением $p_T(t)$. Поскольку подобные вспомогательные величины не отражены в данных, в литературе их принято называть **скрытыми величинами**.

Из выбора совместного распределения зачастую вытекает и выбор метода обучения. Среди возможных методов обучения мы выделим две категории: в случае, когда интерес представляет выбор параметров модели, параметры могут быть получены путем максимизации правдоподобия:

$$\max_{\theta} \log p(\{x^i, y^i\}_{i=1}^n \mid \theta) \tag{9}$$

Если в модели есть скрытые переменные, естественно рассмотреть маргинальное правдоподобие. В литературе подход упоминается как эмпирический байес или методом максимальной обоснованности:

$$\max_{\theta} \log p(\{x^i, y^i\}_{i=1}^n \mid \theta) \tag{10}$$

$$\log p(\{x^i, y^i\}_{i=1}^n \mid \theta) = \log \mathbb{E}_T p(\{x^i, y^i\}_{i=1}^n, T \mid \theta) \tag{11}$$

В случае, когда интерес представляет какая-то из скрытых величин, восстановить их характерные значения можно на основе апостериорного распределения $p(T \mid \{x^i, y^i\}_{i=1}^n, \theta)$. В частности, апостериорное распределение может быть использовано при решении задачи 10. Вычислить апостериорное распределение явно на практике зачастую невозможно, а одним из распространенных подходов к его приближению является вариационный вывод, сводящийся к задаче

$$\max_{\phi} \mathbb{E}_T \log \frac{p(\{x^i, y^i\}_{i=1}^n, T \mid \theta)}{q(T \mid \phi)}, \tag{12}$$

где математическое ожидание берется по случайной величине T с распределением $q(\cdot \mid \phi)$, а оптимизация ведется по параметрам распределения ϕ.

Описанные выше целевые функции можно интерпретировать как суррогатные функции потерь, определенные на основе вероятностного подхода. Поскольку суррогатные функции 9,10,12 никак не зависят от функции потерь $L(\cdot, \cdot)$, в общем случае они могут оказаться несостоятельными. С другой стороны, в сравнении с классической постановкой структурного обучения, вероятностный подход позволяет работать со скрытыми структурными переменными. Это, в свою очередь, дает возможность "сквозного"решения задач, где на каких-то этапах предсказания нужны вспомогательные структурные переменные. Например, при решении дискриминативной задаче в обработке текстов скрытой вспомогательной переменной может быть дерево разбора.

Для описанных выше задач существует ряд общих методов решения. В частных случаях возможно аналитическое решение задач 9,10,12. В общем случае найти аналитическое решение не представляется возможным, а для численного поиска решения но на помощь приходит стохастический градиентный спуск. К стохастической

оптимизации может быть сведена задача 9 в случае, когда параметр θ пробегает дискретное структурное множество. В задачах 10,12 стохастическая оптимизация позволяет оптимизировать математическое ожидание в постановке задачи не прибегая к точному его вычислению. При построении несмещенной оценки градиента математического ожидания, как правило, применяется репараметризационный трюк или метод REINFORCE. Для структурных переменных первый не всегда применим, а второй требует существенных доработок с учетом особенностей выбранной модели.

Ниже дадим описания двух основных подходов к оценке стохастических градиентов. Для задачи вида

$$\max_{\theta} \mathbb{E}_T f(T), \tag{13}$$

где случайная величина T имеет распределение $q(\cdot \mid \theta)$, алгоритм REINFORCE строит несмещенную оценку градиента, преобразовывая выражение для градиента через производную логарифма

$$\nabla_\theta \mathbb{E}_T f(T) = \mathbb{E}_T f(T) \nabla_\theta \log q(T \mid \theta), \tag{14}$$

что позволяет построить несмещенную оценку градиента используя реализацию t случайной величины T:

$$g(t, \theta) = f(t) \nabla_\theta \log q(t \mid \theta). \tag{15}$$

Оценка не накладывает ограничений на вид функции f, но требует эффективный алгоритм генерации t и подсчета $\log q(t \mid \theta)$. Последнее накладывает дополнительные ограничения для некоторых классов дискретных структурных переменных, таких как распределения на основе экспоненциальных семейств. Помимо этого,на практике сходимость алгоритма может быть затруднена высокой дисперсией оценки $g(t, \theta)$, из-за чего возникает необходимость в контрольных переменных для снижения дисперсии.

Репараметризационный трюк позволяет оценить градиенты в задаче 13 в предположении, что случайная величина T может быть представлена в виде $T = h(U, \theta)$ для гладкой f, гладкой по второму аргументу h и некоторой случайной величины U. В соответствии с названием, градиентная оценка получается в результате дифференцирования математического ожидания в новой параметризации

$$\nabla_\theta \mathbb{E}_T f(T) = \nabla_\theta \mathbb{E}_U f(h(U, \theta)) = \mathbb{E}_U \nabla_\theta f(h(U, \theta)), \tag{16}$$

приводя для реализации U к оценке вида

$$g(u, \theta) = \nabla_\theta f(h(u, \theta)) = \left. \frac{\partial f}{\partial t} \right|_{t=h(u,\theta)} \frac{\partial h}{\partial \theta}. \tag{17}$$

В сравнении с оценкой 15, оценка с использование репараметризационного трюка 17 на практике обладает более низкой дисперсией, но накладывает дополнительные ограничения на f и T, явно используя производные. В частности, оценка 17 напрямую не применима для дискретных переменных, позволяя оценивать градиенты лишь для их непрерывных релаксаций.

3 Основные результаты

В данном разделе будут кратко изложены основные результаты, полученные в рамках диссертационного исследования.

3.1 Общие методы

3.1.1 Предсказание порядка на основе вероятностной релаксации

Работа [21] была посвящена методам приближенного вывода в случае, когда структурная скрытая переменная T принимает значения на множестве перестановок. В качестве класса распределений для вывода мы рассмотрели распределение Плакетта-Люса.

Определение 3. Распределением Плакетта-Люса с параметрами $\theta_1, \dots, \theta_n$ называется распределение на перестановках с вероятностью исхода $t \in S_n$ определенной формулой

$$\mathbb{P}_T(T = t; \theta) = \prod_{i=1}^{n} \frac{\exp \theta_{t_i}}{\sum_{j=i}^{n} \exp \theta_{t_j}}. \tag{18}$$

Можно заметить, что определение соответствует выбору n из n элементов без возвращения, где вероятность выбора $i-$ элемента пропорциональна $\exp \theta_i$. Распределение также представляет интерес с точки зрения вероятностной релаксации оптимизационных задач. Ослабим минимум по аргументам функции до минимума среднего значения функции

$$\min_t f(t) \leq \min_\theta \mathbb{E}_T f(T), \tag{19}$$

где T имеет распределение Плакетта-Люса с параметрами θ. Данная оценка гладко зависит от параметров распределения. Также оценку можно сделать сколь угодно точной. Действительно, при шкалировании параметров $\theta' = \theta/\tau$ на стремящуюся к нулю температуру τ распределение стремится к вырожденному. Мода распределения соответствует сортировке параметров в порядке убывания, поскольку такая сортировка доставляет максимум каждого множителя из формулы 18. Поэтому, если сортировка вектора θ совпадает с оптимальной перестановкой τ^*, температурное шкалирование позволит сделать разрыв сколь угодно малым.

Явная формула вероятности и генерация с помощью сэмплирования без возвращения позволяют использовать алгоритм REINFORCE [71] для приближенного вывода в классе распределений Плакетта-Люса. Однако базовая форма алгоритма медленно сходится из-за высокой дисперсии, поэтому в рамках нашей работы мы адаптировали алгоритм RELAX [24] для получения оценок градиента с низкой дисперсией.

Определение 4. Пусть дискретная случайная величина T является функцией $T = H(Z)$ от репараметризуемой случайной величины Z с параметрами θ. Тогда оценка

$$g_{RELAX}(f) = [f(t) - c_\phi(\tilde{z})] \frac{\partial}{\partial \theta} \log \mathbb{P}_T(T = t; \theta) + \frac{\partial}{\partial \theta} c_\phi(z) - \frac{\partial}{\partial \theta} c_\phi(\tilde{z}) \tag{20}$$

для реализации z случайной величины Z, дискретной величины $t = H(z)$ и независимой реализации $\tilde{z}$ условной случайной величины $Z \mid T = t$ является несмещенной оценкой $\mathbb{E}_T f(T)$.

Изначально аналогичная оценка была предложены в [66], где в качестве $c_\phi(\cdot)$ рассматривали функцию f обобщенную на носитель Z. В работе [24] предложили рассматривать дифференцируемую функцию относительно аргумента z и параметров ϕ, настраивая параметры ϕ по мере оптимизации для снижения дисперсии. Обе работы рассматривали случай категориального распределения, в то время как мы обобщили метод на случай распределения Плакетта-Люса.

В основе нашего обобщения лежит эквивалентное определение распределения Плакетта-Люса [73].

Определение 5. Пусть $Z_1, \dots, Z_n$ независимые случайные величины с распределением Гумбеля с соответствующими параметрами $\theta = (\theta_1, \dots, \theta_n)$. Тогда сортировка этих случайных величин T имеет распределением Плакетта-Люса:

$$\mathbb{P}(z_{t_1} \geq \cdots \geq z_{t_n}; \theta) = \prod_{i=1}^{n} \frac{\exp \theta_{t_i}}{\sum_{j=i}^{n} \exp \theta_{t_j}}. \tag{21}$$

Таким образом, случайную величину T можно представить как детерминированную функцию от Z, а для того чтобы воспользоваться оценкой 20 достаточно найти репараметризацию для условного распределения $Z \mid T = t$. В нашей работе предложен алгоритм для репараметризации и эффективной генерации из этого распределения:

Теорема 2. *Рассмотрим независимые в совокупности реализации равномерного распределения $v_1, \dots, v_n \sim U[0,1]$ и реализации распределения Гумбеля $z_1, \dots, z_n$ с параметрами $\theta_1, \dots, \theta_n$. Тогда для перестановки $t = \arg\operatorname{sort}(z_1, \dots, z_n)$ вектор $\tilde{z} = (\tilde{z}_1, \dots, \tilde{z}_n)$ определенный как*

$$\tilde{z}_{t_i} = \begin{cases} -\log(-\log(v_i)) & i = 1 \\ -\log\left(\frac{\log v_i}{\sum_{j=i}^{n} \exp \theta_{t_j}} + \exp(-\tilde{z}_{t_{i-1}})\right) & i \geq 2 \end{cases} \tag{22}$$

является реализацией условного распределения $Z \mid T = t$.

Качество работы предложенного метода вывода мы исследовали на задаче поиска каузальной структуры данных, рассмотрев несколько постановок задачи. Во-первых, мы рассмотрели синтетические данные порожденные на основе модели структурных уравнений (Structured Equation Model, [46]). Для генерации данных мы выбирали случайный направленный ациклический граф $G = (E, V)$ с взвешенной матрицей смежности $W \in \mathbb{R}^{n \times n}$, а затем генерировали данные $X \in \mathbb{R}^{n \times N}$ удовлетворяющие уравнению

$$X = W^T X + \varepsilon, \tag{23}$$

где в качестве ε мы брали гауссовский шум. Это уравнение описывает линейную зависимость, в которой каждая компонента X_i зависит от родителей вершины i в графе G, а также от случайного шума. Задача заключалась в том чтобы по данным X восстановить структуру графа G.

Для сведение задачи к задаче вывода перестановки мы параметризовали искомую матрицу смежности W на основе топологической сортировки: $W = PAP^T$, где $A \in \mathbb{R}^{n \times n}$ была строго верхнетреугольной матрицей, а P была перестановочной матрицей для топологической сортировки графа. Для выбранной параметризации мы решали задачу

$$\min_{P \in \mathbb{P}} \min_{A \in \mathbb{A}} \frac{1}{2N} \|X - PAP^T X\}_F^2 + \lambda \| \operatorname{vec}(A)\|_1 = Q(P, A), \tag{24}$$

где $\mathbb{P}$ - множество перестановочных матриц, а $\mathbb{A}$ - множество строго верхнетреугольных матриц.

Для оптимизации по множество перестановок мы перешли в вероятностной релаксации

$$\min_{\theta} \mathbb{E}_T \min_{A \in \mathbb{A}} Q(P(T), A), \tag{25}$$

сравнивая наш метод с ранее предложенными алгоритмами Gumbel-Sinkhorn [40] и URS [26] на основе релаксации перестановочной матрицы P. В таблице 1 приведены результаты экспериментов для четырех семейств графов с 20 вершинами. Предложенные для сравнения алгоритмы существенно уступают нашему подходу как и по качестве оптимизации целевой функции, так и по структурным метрикам SHD, SHD-CPDAG и SID. Мы также доработали алгоритмы Sinkhorn и URS, добавив дополнительные ограничения при оптимизации, получив сравнимые результаты. Полное описание этого эксперимента, а также остальные эксперименты можно найти в приложенной к реферату статье.

3.1.2 Гарантии обучения при использовании квадратичных суррогатных функций

В работе [61] мы провели анализ суррогатных функций потерь в задаче структурного предсказания. Основной теоретический результат работы состоит в усиленной верхней оценке на калибровочную функцию квадратичной суррогатной функции, позволяющий получить нетривиальные гарантии в случае когда суррогатная функция потерь не является состоятельной.

Следуя введенным в разделе 2.2 обозначениям, введем квадратичную суррогатную функцию потерь

$$\Phi_{quad}(f, y) := \frac{1}{2k}\|f + L(:, y)\|_2^2 = \frac{1}{2k}\sum_{\hat{y}\in\mathcal{Y}}\left(f_{\hat{y}}^2 + 2f_{\hat{y}}L(\hat{y}, y) + L(\hat{y}, y)^2 + L(\hat{y}, y)^2\right). \quad (26)$$

Как было отмечено выше, предсказание f зачастую переметризуют с помощью дополнительной матрицы $F : f(x) = Fg(x)$. Ранее в [44] была полученная оценка на квадратичную функцию потерь в предположении, что линейная оболочка $\mathcal{F}$ столбцов матрицы F совпадает с линейной оболочкой столбцов функции потерь.

Теорема 3. *Для любой матрицы потерь L, соответствующего квадратичного суррогата Φ_{quad} и пространства предсказаний $\mathcal{F}$, содержащего столбцы матрицы L, калибровочная функция имеет нижнюю оценку*

$$H_{\Phi_{quad},L,\mathcal{F}}(\varepsilon) \geq \frac{\varepsilon^2}{2k\max_{i\neq j}\|P_{\mathcal{F}}\Delta_{ij}\|_2^2} \geq \frac{\varepsilon^2}{4k}, \quad (27)$$

для оператора $P_{\mathcal{F}}$ ортогональной проекции на $\mathcal{F}$ и вектора $\Delta_{ij} = e_i - e_j \in \mathbb{R}^k$, где e_c обозначает c-й стандартный базисный вектор в $\mathbb{R}^k$.

Второе неравенство тривиально и приводит к оценке, полученной в более ранней работе [14]. С другой стороны, по мере уменьшения $\mathcal{F} \subsetneq \mathbb{R}^k$ норма проекции $\|P_{\mathcal{F}}\Delta_{ij}\|_2^2$ падает, приводя в более точным нижним оценкам калибровочной функции. Минимальное удовлетворяющее условиям теоремы множество предсказаний будет $\mathcal{F} = \operatorname{span} L$.

В нашей работе мы ослабили ограничение, требующее $\operatorname{span} L \subset \mathcal{F}$, получив следующую общую оценку.

Теорема 4. *Для любой марицы потерь L, соответствующего квадратичного суррогата Φ_{quad} и пространства предсказаний $\mathcal{F}$ калибровочная функция имеет нижнюю оценку*

$$H_{\Phi_{quad},L,\mathcal{F}}(\varepsilon) \geq \min_{i\neq j}\max_{v\geq 0}\frac{(\varepsilon v - \xi_{ij}(v))_+^2}{2k\|P_{\mathcal{F}}\Delta_{ij}\|_2^2}, \text{ где } \xi_{ij}(v) := \|L^T(vI_k - P_{\mathcal{F}})\Delta_{ij}\|_\infty, \quad (28)$$

ER1				
	Val Q - Q*	SHD	SHD-CPDAG	SID
PL-RELAX	15.7±27.3	14.4±5.3	16.0±6.2	61.0±48.7
SINKHORN_{ECP}	10.4±8.7	15.8±4.7	17.0±6.0	84.8±56.3
URS_{ECP}	27.5±34.2	20.6±6.3	21.4±7.2	96.8±74.6
SINKHORN	1651.2±3050.4	24.0±6.1	25.0±6.7	131.2±76.5
GREEDY-SP	N/A	18.6±13.5	18.0±16.6	74.0±53.5
RANDOM	895.1±1270.3	37.8±5.2	38.8±4.9	146.8±79.9
SF1				
	Val Q - Q*	SHD	SHD-CPDAG	SID
PL-RELAX	-1.5±0.2	4.0±0.6	4.6±0.5	4.2±0.7
SINKHORN_{ECP}	1.9±4.3	6.6±2.2	6.6±2.4	10.4±5.0
URS_{ECP}	3.0±2.0	10.6±2.0	10.6±1.6	14.4±4.0
SINKHORN	38.3±26.2	19.0±0.0	19.0±0.0	35.0±2.4
URS	38.3±26.2	19.0±0.0	19.0±0.0	35.0±2.4
GREEDY-SP	N/A	2.0±1.4	0.0±0.0	7.0±5.1
RANDOM	94.0±36.4	36.2±2.6	36.6±2.3	48.6±14.7
ER4				
	Val Q - Q*	SHD	SHD-CPDAG	SID
PL-RELAX	468.8±208.4	71.0±5.9	72.6±3.9	289.6±9.1
SINKHORN_{ECP}	2519.0±3715.2	78.0±6.1	78.8±5.5	302.2±15.8
URS_{ECP}	1011.4±745.5	75.8±2.9	76.6±2.9	300.2±20.3
SINKHORN	126284.6±194386.3	88.8±6.0	91.0±5.7	330.0±14.1
GREEDY-SP	N/A	103.4±10.9	105.6±10.5	288.6±14.7
RANDOM	109891.2±74968.7	113.0±4.9	114.4±4.1	330.6±9.2
SF4				
	Val Q - Q	SHD	SHD-CPDAG	SID
PL-RELAX	-5.8±1.2	20.0±4.3	20.0±4.1	48.4±16.2
SINKHORN_{ECP}	-0.4±2.4	25.6±5.6	25.8±5.9	58.6±19.7
URS_{ECP}	8.5±11.8	30.2±5.8	30.6±5.2	72.2±25.0
SINKHORN	158.2±99.9	44.6±5.8	44.8±6.1	103.6±20.8
URS	140.7±140.6	42.0±5.4	42.8±5.1	89.8±20.4
GREEDY-SP	N/A	50.6±31.5	49.8±32.3	69.0±43.2
RANDOM	635.5±182.6	98.2±6.1	99.2±5.5	168.8±29.6

Таблица 1: Метрики для графов из 20 вершин

оператор $P_{\mathcal{F}}$ задает ортогональную проекцию на $\mathcal{F}$, функция $(x)_+^2 := [x > 0]x^2$ задает правую ветвь параболы и $\Delta_{ij} := e_i - e_j \in \mathbb{R}^k$, где e_c обозначает c-й стандартный базисный вектор в $\mathbb{R}^k$.

Полагая $v = 1$ во введенной выше оценке, мы также можем получить упрощенное выражение

$$H_{\Phi_{quad},L,\mathcal{F}}(\varepsilon) \geq \min_{i \neq j} \frac{(\varepsilon - \xi_{ij})_+^2}{2k\|P_{\mathcal{F}}\Delta_{ij}\|_2^2}, \text{ где } \xi_{ij} := \|L^T(I_k - P_{\mathcal{F}})\Delta_{ij}\|_\infty. \tag{29}$$

Можно заметить, что при $\operatorname{span} L \subset \mathcal{F}$ новая оценка 28 не хуже старой 27. Действительно, выражение внутри совпадает со старой оценкой при $v = 1$, но может принимать даже больший значения при $v \neq 1$. В случае, когда $\mathcal{F}$ не содержит столбцы $\mathcal{F}$ нижняя будет огибающей семейства кривых с параметром v. Каждая из кривых является правой ветвью параболы, смещенной вправо. Вблизи нуля нижняя оценка равна нулю из-за несостоятельности суррогата. Точка отрыва равна $\eta = \frac{\xi_{ij}(v)}{v}$ и определяет уровень состоятельности суррогата в широком смысле.

Помимо вывода общей оценки, мы вычисляем константы в неравенстве и проводим анализ нескольких популярных функций потерь. Приведем в качестве иллюстрации функцию потерь mAP (mean average precision), применяемую в задачах ранжирования [10, 9, 49]. В данном случае предсказание модели $\sigma \in \hat{\mathcal{Y}} = S_r$ является перестановкой на r элементах, а метки $y \in \mathcal{Y} = \{0, 1\}^r$ представляют из себя бинарные вектора длины r. Функция потерь $L_{mAP}(\sigma, y)$ усредняет точность ранжирования для разных уровней полноты:

$$L_{mAP}(\sigma, y) := 1 - \frac{1}{|y|} \sum_{p:y_o=1}^{r} \frac{1}{\sigma(p)} \sum_{q=1}^{\sigma(p)} y_{\sigma^{-1}(q)} = 1 - \sum_{p=1}^{r} \sum_{q=1}^{p} \frac{1}{\max(\sigma(p), \sigma(q))} \frac{y_p y_q}{|y|}. \tag{30}$$

Выше норма вектора равна $|y| = \sum_{p=1}^r y_p$. Второе выражении для матрицы потерь приводит к двум естественным определениям $\mathcal{F}$, которые приведены ниже. Для первой параметризации зададим $\mathcal{F}_{mAP} = \operatorname{span} F_{mAP}$ через линейную оболочку столбцов матрицы $F_{mAP} \in \mathbb{R}^{r! \times \frac{1}{2}r(r+1)}$ с элементами $(F_{mAP})_{\sigma,pq} := \frac{1}{\max(\sigma(p),\sigma(q))}$. Из определения L_{mAP} следует, что $\operatorname{span} L_{mAP} = \operatorname{span} F_{mAP}$, а квадратичная суррогатная функция потерь будет состоятельной. С другой стороны, вывод к этой модели сводится к целочисленной задаче квадратичного программирования $\max_{\sigma \in S_r}(F_{mAP}\theta)_\sigma$, которая с общем случае не имеет эффективного алгоритма решения. Для второй параметризации зададим $\mathcal{F}_{sort} = \operatorname{span} F_{sort}$ как линейную оболочку матрицы $F_{sort} \in \mathbb{R}^{r! \times R}$ с элементами $(F_{sort})_{\sigma,p} := \frac{1}{\sigma(p)}$. Предсказание $\max_{\sigma \in S_r}(F_{sort}\theta)_\sigma$ в этой параметризации сводится к сортировке элементов вектора θ, что делает вторую параметризацию предпочтительнее первой. С другой стороны, $\mathcal{F}_{sort}$ не содержит столбцы матрицы L_{mAP}, из-за чего квадратичный суррогат оказывается несостоятельным.

На рисунке 1 приведены графики описанных выше оценок для функции потерь L_{mAP}. Из-за несостоятельности суррогатной функции потерь, график для F_{sort} отрывается от нуля лишь при $\varepsilon > 0$. В то же время, начиная с некоторых значений ε, калибровочная функция для F_{sort} оказывается выше калибровочной функции F_{mAP}. На практике это означает, что для ограниченного уровня точности параметризация F_{sort} позволяет получить лучшие гарантии обучения нежели F_{mAP}, позволяя также эффективно строить предсказания.

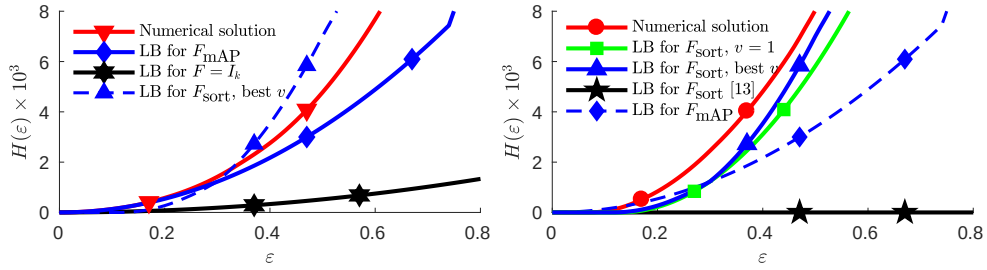

Рис. 1: Слева: состоятельная калибровочная функция для F_{mAP}; справа: несостоятельная калибровочная функция для F_{sort}

3.2 Приложения

3.2.1 Предсказание ядер сверточной нейронной сети

Работа [1] предлагает интерпретировать параметры сверточной нейронной сети как структурную скрытую переменную. По сравнению с базовым моделями байесовских нейронных сетей, структурное априорное распределение параметров сети учитывает зависимости между отдельными весами сверточных фильтров. В экспериментах предложенная модификация повысила качество классификации в условиях ограниченной обучающей выборки, позволила ускорить обучение сети, позволила выделять низкоразмерные представления данных без дополнительного обучения.

Байесовская нейронная сеть - дискриминативная модель на стыке байесовский методов машинного обучения и глубинного обучения. Совместное распределение

$$p(y^1, \dots, y^n, \theta \mid x^1, \dots x^n) = \left[\prod_{i=1}^{n} p(y^i \mid x^i, \theta)\right] p(\theta) \tag{31}$$

на метки классов $y^1, \dots, y^n$ и веса модели θ, как правило, состоит из набора независимых распределений на каждый отдельный вес $p(\theta) = \prod_j p(\theta_j)$ и правдоподобия метки $p(y^i \mid x^i, \theta)$ при условии входного объекта x^i и весов θ. Предсказание в рамках этой модели подразумевает вычисление апостериорного распределения на веса $p(\theta \mid \{x^i, y^i\}_{i=1}^n)$, на основе которого затем можно вычислить распределение меток

$$p(y_{test} \mid \{x^i, y^i\}_{i=1}^n, x_{test}) = \int p(y_{test} \mid x_{test}, \theta) p(\theta \mid \{x^n, y^n\}_{n=1}^N) \mathrm{d}\theta \tag{32}$$

. На практике апостериорное распределение ищут в приближенном виде с помощью вариационного вывода, решая задачу

$$\max_{\phi} \left[\mathbb{E}_{\Theta} \log p(y^1, \dots, y^n \mid \Theta, x^1, \dots, x^n) - KL(q(\Theta; \phi) || p(\Theta))\right], \tag{33}$$

где математическое ожидание берется по вариационному распределению с плотностью $q(\theta; \phi)$ и параметрами ϕ. Предположение о независимости параметров в априорном распределении упрощает параметризацию модели. С другой стороны, веса сверточных фильтров обученной нейронной сети не ведут себя как независимые случайные величины. Качественно, веса плавно меняются в зависимости от расположения в фильтре (пример приведен на графике 2. Более того, в приложениях обученные параметры сверточной сети могут быть использованы в новой задаче на схожем домене [74, 57]. Например, в случае изображений, обученные на ImageNet сверточные сети могут быть адаптированы к другим задачам компьютерного зрения.

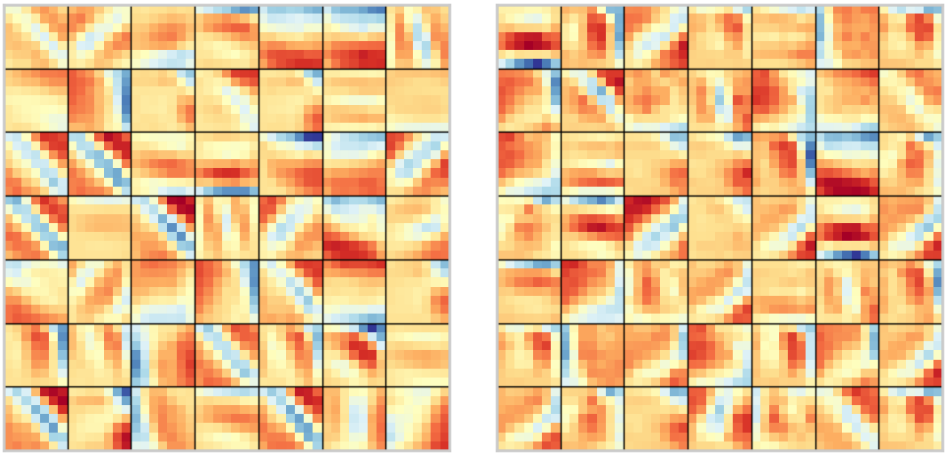

Рис. 2: Слева: фильтры обученной сверточной сети. Справа: фильтры, полученные из приближения.

Статья предлагает рассмотреть распределение сверточных фильтров на некотором домене. Чтобы его приблизить, мы предлагаем выбрать базовую задачу для данного домена и обучить несколько сверточных сетей. Базовая задача должна быть репрезентативна для данного домена: обучающие примеры должны быть разнообразны, а выучиваемые сетью представления должны быть достаточно информативны. В работе мы рассматривали задачи классификации изображений, рассматривая задачи новыми наборами классов. Обучив несколько сверточных сетей, мы можем использовать их фильтры в качестве приближения распределения фильтров. При работе с таким приближением возникает две проблемы. Во-первых, для работы с распределением необходимо хранить множество сверточных сетей в памяти. Во-вторых, для построенного приближения неизвестна плотность, необходимая для вычисления целевой функции байесовской нейронной сети. Поэтому мы предлагаем приблизить распределение фильтров с помощью вспомогательной порождающей модели на основе вариационного авто-кодировщика.

При обучении мы предлагаем заменить априорное распределение $p(W)$ на оценку, полученную на основе вариационного авто-кодировщика. Таким образом, мы приходим к нижней оценке на маргинальное правдоподобие

$$\log p(\{y^i\}_{i=1}^n \mid \{x^i\}_{i=1}^n) \geq \mathbb{E}_\Theta \left[\log p(\{y^i\}_{i=1}^n \mid \Theta, \{x^i\}_{i=1}^n)\right. \tag{34}$$

$$+ \mathbb{E}_Z \log \frac{p(\Theta \mid Z; \chi) p(Z)}{r(Z \mid \Theta; \psi)} \tag{35}$$

$$\left. - \log q(\Theta; \phi)\right], \tag{36}$$

где $q(\Theta; \phi)$ является вариационным приближением параметров сети, распределения $p(\Theta \mid Z; \chi)$ и $r(Z \mid \Theta; \psi)$ определены вариационным авто-кодировщиком, а математическое ожидание по случайному вектору Z вычисляется относительно распределения $r(Z \mid \Theta; \psi)$ При обучении байесовской нейронной сети мы будем использовать эту оценку в качестве целевой функции. Первое слагаемое в оценке соответствует стандартной кросс-энтропийной функции потерь, а второе слагаемое поощряет близость приближенного апостериорного распределения сети $q(\Theta; \phi)$ к подобранному эмпирически априорному распределению параметров сверточной сети для данного домена $p(\Theta)$.

Для оценки предложенного подхода мы провели ряд экспериментов, оценивающий

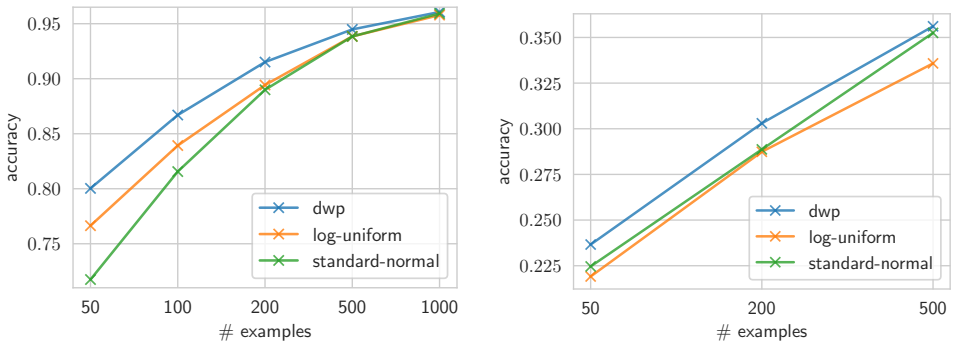

Рис. 3: Графики качества классификации в зависимости от размера обучающей выборки. Слева: MNIST, справа: CIFAR-10.

способность к обучению в условиях ограниченной выборки, качество представлений полученных из сети после инициализации и скорость обучения в зависимости от выбора априорного распределения для инициализации сети. Здесь мы ограничимся описанием первого эксперимента, подробное описание остальных можно найти в оригинальной работе.

Изучая обучение байесовской сети с ограниченной выборкой, мы рассмотрели задачу классификации на данных MNIST и CIFAR-10. В качестве модели мы взяли сверточные сети, состоящие из нескольких сверточных слоев вместе с несколькими полносвязными слоями на выходе. Для сверточных слоев мы обучили априорное распределение на данных NotMNIST и CIFAR-100, а на полносвязные слои мы обучали без применения вариационного вывода. Для сравнения мы рассмотрели распространенные в литературе априорные распределения для параметров сети: гауссовское распределение и log-uniform распределение, гарантирующие инвариантность априорного распределения к масштабу параметров.

Варьируя размер обучающий выборки, мы получили три вида классификаторов. Как показывает график 3, сеть с предложенным априорным распределением показывает лучшие результаты. На данных MNIST разница в качестве пропадает при достаточном размере обучающей выборки. На данных CIFAR-10 качество равномерно выше. Мы предполагаем, что разницу можно объяснить простотой задачи классификации на данных MNIST: тысячи примером оказывается достаточно чтобы извлечь необходимую информацию из данных.

3.2.2 Предсказание конфигурации многопользовательского канала связи

Описание вероятностной модели. В работе [60] мы рассмотрели задачу оценки параметров многопользовательского канала связи. Для установления соединения по выделенным частотам пользователи посылают специальные кодовые сигналы, которые затем принимает и обрабатывает станция сотовой связи. Канал связи многопользовательский, поэтому некоторые частоты могут занимать несколько пользователей и система должна уметь детектировать пользователей получая суперпозицию присланных кодовых сигналов. На практике для упрощения задачи предполагается, что пользователей достаточно мало.

Фактически задачу можно интерпретировать как задачу структурного предсказания, где на основе полученных сигналов необходимо выбрать разреженный бинарный вектор с блочной структурой. Стандартные методы решения основаны на модифика-

циях алгоритмов compressed sensing. В работе [70] работу был предложен подход на основе байесовской линейной регрессии. Баейсовская линейная регрессия позволяет находить разреженные решения линейных систем уравнений, что требуется в данной задаче. В нашей работе мы адаптировали стандартную модель байесовской линейной регрессии к специфике задачи, учтя блочную структуру искомого решения, а также предложили более быстрый алгоритм решения системы.

Рассмотренная нами модель может быть описана системой линейных уравнений вида

$$y = \kappa\theta + z, \tag{37}$$

где вектор y соответствует принятому сигналу, матрица κ фиксирована и задана протоколом связи, вектор θ неизвестен и описывает конфигурацию канала, а z моделирует возникающий при передаче сигнала шум. В качестве модели шума мы использовали гауссовское распределение с заранее известной дисперсией ρ. Помимо этого, вектор θ имеет блочную структуру

$$\theta = (c_{11}t_1, \ldots, c_{1Q}t_1, c_{21}t_2, \ldots, c_{2Q}t_2, \ldots, c_{N1}t_N, \ldots, c_{NQ}t_N), \tag{38}$$

где бинарные переменные $t_1, \ldots, t_N \in \{0, 1\}$ равны единице, если пользователь активен, а величины $c_{11}, \ldots, c_{NQ} \in \mathbb{R}$ варьируются в зависимости от параметров канала связи. В рамках этой модели, мы в первую очередь заинтересованы в восстановлении величин $t_1, \ldots, t_N$, которые указывают на активных пользователей в канале. Помимо этого, интерес представляет оценка вектора w, поскольку она содержит дополнительные параметры канала связи такие как уровень затухания сигнала.

Для решения задачи мы рассмотрели модель байесовской линейной регрессии, заданную вероятностным распределением

$$p(y, \theta; \rho, \gamma) = p(y \mid \theta; \rho)p(\theta; \gamma) \tag{39}$$

$$p(Y = y \mid \Theta = \theta; \rho) = \mathcal{N}(y \mid \kappa\theta; \rho I) \tag{40}$$

$$p(\Theta = \theta; \gamma) = \mathcal{N}(\theta \mid 0, \operatorname{diag}(\underbrace{\gamma_1, \ldots, \gamma_1}_{Q}, \ldots, \underbrace{\gamma_N, \ldots, \gamma_N}_{Q})). \tag{41}$$

Плотность $p(Y = y \mid \Theta = \theta; \rho)$ выше задает правдоподобие наблюдений, а $p(\Theta = \theta; \gamma)$ задает априорное распределение блочной структуры. Отметим, что в предложенных ранее работах использовалась базовая модель регрессии, которая не учитывала блочную структуру априорного распределения, в отличие от рассмотренной нами модели.

Чтобы оценить конфигурацию канала мы максимизировали обоснованность $p(y; \rho, \gamma)$ по параметрам априорного распределения γ. Также как и [70], для вывода мы использовали EM-алгоритм, поочередно оценивая апостериорное распределение $p(\theta \mid y, \rho, \gamma)$ на E-шаге и максимизируя оценку на обоснованность по параметрам γ на M-шаге. Для оптимизации на M-шаге мы воспользовались итеративной схемой, предложенной в [64], которая улучшила скорость вывода в модельных экспериментах по сравнению с раннее рассмотренными схемами для нашей задачи [70].

Результаты симуляции. Мы провели симуляцию для оценки качества работы предложенной схемы. Мы использовали Релеевскую модель затухания для моделирования амплитуды сигнала, рассмотрели канал с 6 активными пользователями из $N = 36$, каждый из которых использовал $Q = 5$ частот. В качестве кодовых последовательностей в матрице κ мы рассмотрели последовательности Задова-Чу длины 20. На графике 4 приведена зависимость средней доли неправильно определенных пользователей $\text{UDER} = \mathbb{E}\left[\frac{\sum_{n=1}^{N}[\hat{a}_n \neq a_n}{N}\right]$ от количества итераций EM-алгоритма для

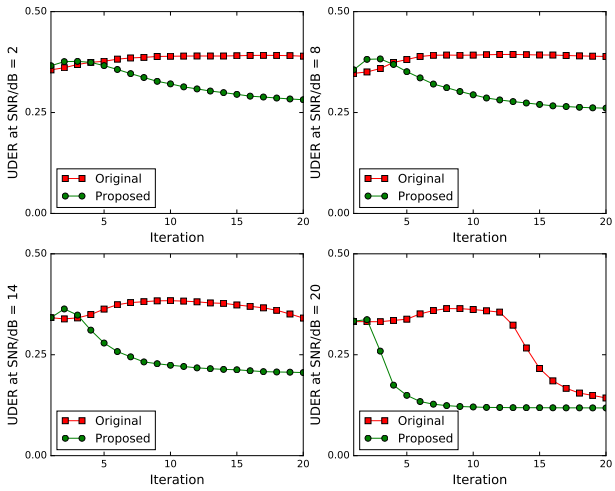

Рис. 4: Зависимость ошибки детекции пользователей UDER от количества итераций EM-алгоритма

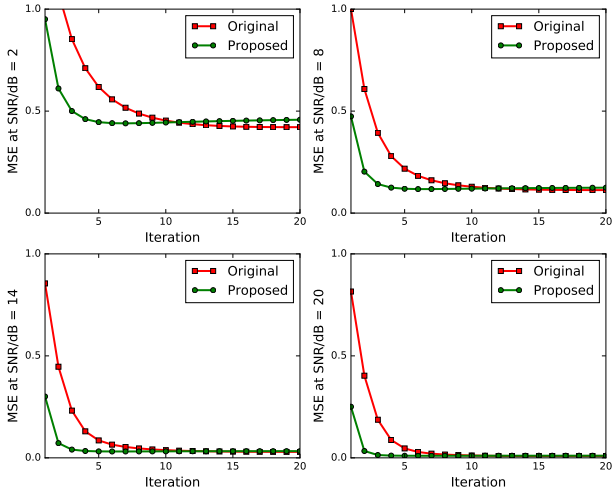

Рис. 5: Зависимость ошибки оценки параметров канала связи Θ от количества итераций EM-алгоритма

разных уровней отношения сигнал-шум в канале связи. Во всех четыре случаях предложенная схема сходится быстрее, чем оригинальная схема. Более того, для высоких уровней шума оригинальная схема в среднем не сходится. На графике 5 приведена зависимость средней ошибки оценки Θ от количества итераций EM-алгоритма. Также как и в предыдущем эксперименте, предложенная схема показывает лучшую сходимость. Примечательно, что по метрике MSE оригинальная схема достигает сравнимых результатов даже для высоких уровней шума.

3.2.3 Пред-обработка данных геологической разведки с использованием скрытых марковских цепей

Описание вероятностной модели. В работе посвященной анализу данных геологической разведки [59] мы адаптировали скрытую марковскую цепь к задаче предобработки и восстановления пропусков в данных геологической разведки. Скрытая марковская цепь является одной из классических вероятностных моделей со скрытой структурной переменной: скрытая переменная задана марковской цепью с дискретными состояниями, наблюдения независимы при условии скрытой переменной, а для

настройки параметров и вывода на скрытую переменную применяется EM-алгоритм.

На первых этапах, геологическая разведка строит модель месторождения. Для этого на территории месторождения бурят множество скважин, в каждую из них опускают ряд датчиков. По мере продвижения вглубь датчики считывают различные физические характеристики скважины в зависимости от глубины. Полученные данные образуют последовательности, именуемые в рамках области "каротажами". Затем, на основе этих данных эксперт-петрофизик отмечает участки глубины скважин, представляющие интерес с точки зрения добычи нефти. Для построения итоговой разметки эксперт также проводит выравнивание данных, дополнительную калибровку и поиск аномалий.

Задача данного исследования заключалась к автоматизации работы эксперта-петрофизика. Накопленные за долгие годы результаты работы экспертов позволяют свести разметку скважин к задаче обучения с учителем, однако предсказания экспертов носят субъективный характер и могут давать недостаточно надежный обучающий сигнал. Поэтому мы рассмотрели модель обучения без учителя, которая могла бы позволять получить согласованные предсказания для всего месторождения.

Опишем предложенную вероятностную модель. Мы предполагаем, что нам даны каротажи $x^1, \ldots, x^K$ для K скважин, $x^k \in \mathbb{R}^{l_k \times d}$. Для каждой скважины на данном уровне глубины показания прибора определяется характеристиками почвы. Мы предположили, что характеристики почвы можно описать последовательностью из m состояний марковской цепи. Для этого мы вводим скрытые случайные вектора $T^1, \ldots, T^K$, $T^k_l \in \{1, \ldots, m\}, l = 1, \ldots, L^k k$, начальное распределение $P(T^k_1 = t; \pi) \propto \pi_k, \pi \in \mathbb{R}^m_+$ и парные распределения $P(T^k_l = t \mid T^k_{l-1} = s; \tau) \propto \tau_{ts}, \tau \in \mathbb{R}^{m^2}_+$. Совместная зависимость элементов цепи также позволяет поощрять одинаковые состояния у соседних величин. Продолжительные участки цепи с одинаковыми состояниями соответствуют однородным участкам скважины, вдоль которых характеристики почвы не изменяются. На практике характеристики почвы нам неизвестны, поэтому марковская цепь задает скрытую переменную в модели. Но нам известны показания приборов, записанные в каротажах. Мы предположили, что для каждого типа почвы показания приборов следуют многомерному нормальному закону $p(x^k_l \mid T^k_l = t; \mu, \Sigma) = \mathcal{N}(x^k_l \mid \mu_t, \Sigma_t), \mu \in \mathbb{R}^{m \times d}, \Sigma \in \mathbb{R}^{m \times d^2}$.

Также известно, что на показания приборов влияет калибровка прибора перед записью. Предполагая, что калибровка задает линейное преобразование показаний $x^k_l = \alpha_k \odot \hat{x}^k_l + \beta_k$ для данного наблюдения x^k_l и откалиброванного наблюдения $\hat{x}^k_l$, для учета этой особенности для каждой скважины мы ввели дополнительные параметры калибровки $\alpha \in \mathbb{R}^{K \times d}_*, \beta \in \mathbb{R}^{K \times d}$. Для параметров $\Theta = (\pi, \tau, \alpha, \beta)$ итоговая модель наблюдений имела вид

$$p(x^i, t^i{}_{i=1}^K; \Theta) = \tag{42}$$

$$\prod_{k=1}^{K} [(\pi_{t^k_1} \prod_{l=2}^{L_k} \tau_{t^k_l t^k_{l-1}}) \times \tag{43}$$

$$\prod_{l=1}^{L_k} \mathcal{N}(x^k_l \mid \alpha_k \odot \mu_{t^k_l} + \beta_k, \operatorname{diag}(\alpha_k) \Sigma_{t^k_l} \operatorname{diag}(\alpha_k))]. \tag{44}$$

Для настройки модели мы использовали алгоритма Баума-Велча, максимизируя нижнюю оценку на обоснованность модели. Вычислив оценку, мы оптимизировали параметры с помощью градиентного спуска, подбирая начальные приближения для лучшей сходимости. Поскольку наблюдения в модели следуют нормальному закону,

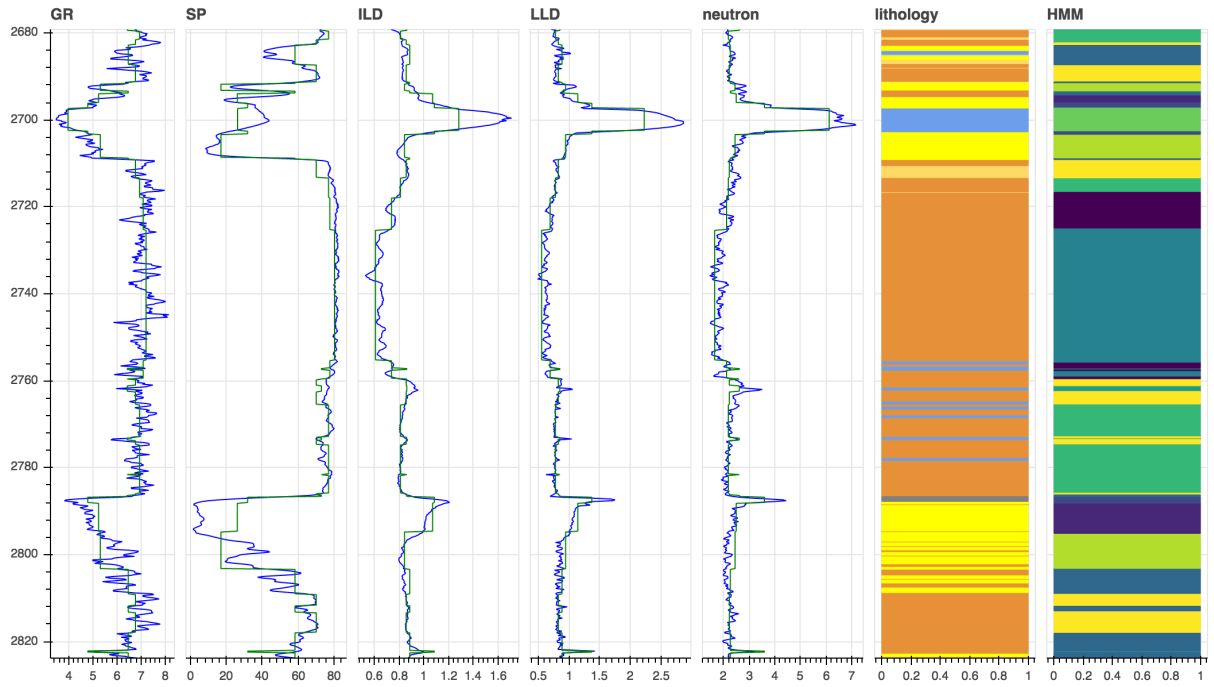

Рис. 6: Иллюстрация работы модели для одной скважины. Слева: пример наблюдаемых каротажей (синий) и их приближения (зеленый); справа: пример данной разметки и найденных состояний марковской цепи

мы могли работать с данными с пропусками, используя в качестве правдоподобия маргинальные распределения без учета пропущенных данных. Для вывода скрытых состояний цепи мы использовали алгоритм Витерби. Ниже приведем результаты работы модели на синтетических месторождениях, а также на Приобском месторождении [3].

Эмпирические результаты. В рамках синтетического месторождения нам даны измерения для скважин, а также достоверно известны типы почв. Благодаря этому мы смогли качественно сопоставить скрытые состояния марковской цепи с истинными породам почв, исключив фактор субъективной интерпретации данных экспертом. Слева график 6 содержит каротажи (синие линии), а также предсказания построенной нами модели (зелёная линия). В то время как нам удалось достаточно точно повторить поведение каротажей, мы не получили взаимно-однозначного соответствия между скрытыми состояниями и типами почв. Разметка типов почв и скрытых состояний скважины приведена на графике 6 справа. Выбранное число скрытых состояний превышало число типов почв: повышение числа скрытых состояний улучшает приближение каротажей, но делает скрытые состояния менее интерпретируемыми. На графике 7 мы привели соответствие между скрытыми состояниями и типами почв по всему месторождению. Большая часть скрытых состояний соответствует аргиллиту, преобладающему в месторождении. Модели также удалось отделить плотные породы и песчаники, но ни одно из скрытых состояний не соответствует алевролиту.

Затем мы применили модель для пред-обработки данных на Приобском месторождении. Базовая модель предсказывала коллекторные слои (слои, представляющие интерес с точки зрения добычи нефти) с помощью бинарной классификации на основе рекуррентной нейронной сети [3]. В базовой модели показания приборов стандартизовали для учета раскалибровки. Мы, в свою очередь, использовали по-

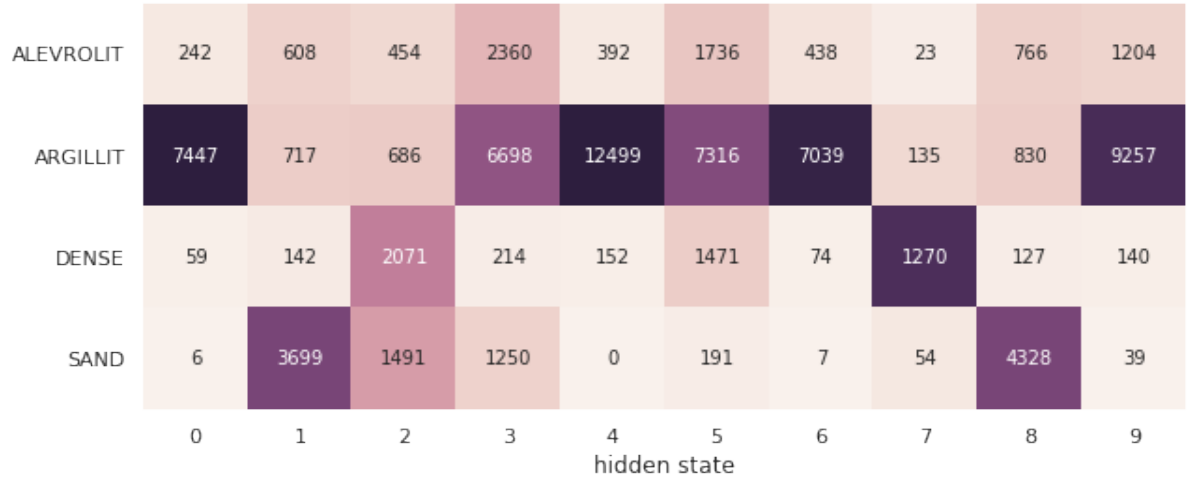

Рис. 7: Соответствие между типами почв и найденными состояниями марковской цепи

правку на предсказанные моделью параметры калибровки α, β. Смена алгоритма пред-обработки не дала значимого улучшения качества предсказания, повысив F1 меру с 0.72 до 0.74.

Наконец, мы использовали модель для восстановления пропусков в данных. Мы рассмотрели тестовые скважины, для которых в выборке нет значений ILD (расшифровка) и LLD (расшифровка) каротажей. Затем мы сравнили две стратегии восстановления пропусков: замена значения каротажа средним по месторождению, а также наш подход, заключавшийся в восстановлении значений каротажа по остальным каротажам с использованием марковской цепи. Предложенное решение позволило повысить качество предсказания для рассмотренных скважин с F1=0.37 до F1=0.56. Таким образом, предложенный нами подход позволяет улучшать качество нахождение коллекторных слоев благодаря совместной калибровке и восстановлению пропусков в данных.

4 Заключение

Описанные выше результаты покрывают различные аспекты задач структурного предсказания, включая теоретический анализ классической постановки задачи, работа со скрытыми структурными переменными на основе вероятностного подхода, а также приложения описанных решений к реальным задачам. В завершение кратко резюмируем представленные результаты.

1. Предложен метод оптимизации по перестановкам на основе вероятностной релаксации и алгоритма REINFORCE, для улучшения сходимости метода разработаны контрольные переменные. Работа метода продемонстрирована в задаче выявления каузальных связей в данных, где в качестве структурной переменной выступает топологическая сортировка направленного ациклического графа связей. Предложенный метод позволил существенно улучшить метрики качества восстановления структуры связей в сравнении с аналогичными методами градиентной оптимизации. Поскольку рассмотренный метод оптимизации не вводит дополнительных предположений о целевой функции и фактически является методом оптимизации нулевого порядка, в будущем он также может найти приме-

нение при прямой оптимизации целевой функции в задачах структурного предсказания (без использования вспомогательных суррогатных функций потерь), а также при амортизированном выводе перестановок.

2. Для задачи обучения с учителем со структурной переменной был проведен анализ общего подхода к обучению, основанного на квадратичных суррогатных функциях потерь. Был рассмотрен случай несостоятельных суррогатных функций потерь, для которого удалось получить гарантии точности оптимизации ожидаемого риска. Предполагая ограниченную точность оптимизации и фиксированное число объектов обучающей выборки, проведенный анализ позволил получить более сильные гарантии относительно значений ожидаемого риска. С практической точки зрения, рассмотренная постановка также приводит к более эффективным алгоритмам вывода.

3. Разработан ряд приложений на основе вероятностного подхода к структурному предсказанию. Модель вариационного авто-кодировщика адаптирована для оценки параметров сверточной нейронной сети на основе априорных знаний о распределении совокупности параметров для данного домена. В данном случае, в параметры сверточных фильтров выступают в качестве скрытой структурной переменной, а предложенный подход позволяет повысить точность предсказания байесовских нейронных сетей на смежные домены. В задаче оценки параметров многопользовательского канала связи в качестве скрытой структурной переменной выступает подмножество активных пользователей. В данном случае для оценки структурной переменной была предложена улучшенная вероятностная модель, а также ускорен алгоритм вывода. Наконец, в задаче интерпретации данных геофизической разведки была предложена вероятностная модель на основе скрытых марковских цепей. Предложенная модель сопоставляет структурные переменные, в данном случае скрытые состояния цепи, физическим характеристикам скважин и моделирует совместное распределение этих характеристик. На основе восстановленных скрытых состояний, предложен подход к восстановлению пропусков в данных, а также детекции аномалий.

Список литературы

[1] Andrei Atanov, Arsenii Ashukha, Kirill Struminsky, Dmitriy Vetrov, and Max Welling. The deep weight prior. In *International Conference on Learning Representations*, 2018.

[2] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[3] Boris Belozerov, Nikita Bukhanov, Dmitry Egorov, Adel Zakirov, Oksana Osmonalieva, Maria Golitsyna, Alexander Reshytko, Artyom Semenikhin, Evgeny Shindin, and Vladimir Lipets. Automatic well log analysis across priobskoe field using machine learning methods. In *SPE Russian Petroleum Technology Conference*. OnePetro, 2018.

[4] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[5] Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable pertubed optimizers. *Advances in neural information processing systems*, 33:9508–9519, 2020.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[7] Céline Brouard, Marie Szafranski, and Florence d'Alché Buc. Input output kernel regression: Supervised and semi-supervised structured output prediction with operator-valued kernels. *Journal of Machine Learning Research*, 17:np, 2016.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[9] David Buffoni, Clément Calauzenes, Patrick Gallinari, and Nicolas Usunier. Learning scoring functions with order-preserving losses and standardized supervision. In *ICML*, 2011.

[10] Clément Calauzènes, Nicolas Usunier, and Patrick Gallinari. On the (non-) existence of convex, calibrated surrogate losses for ranking. In *NIPS*, 2012.

[11] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892. PMLR, 2018.

[12] Justin Chiu and Alexander M Rush. Scaling hidden markov language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1341–1349, 2020.

[13] Anna Choromanska, Alekh Agarwal, and John Langford. Extreme multi class classification. In *NIPS Workshop: eXtreme Classification, submitted*, volume 1, pages 2–1, 2013.

[14] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. *Advances in neural information processing systems*, 29:4412–4420, 2016.

[15] Shay B Cohen and Noah A Smith. Joint morphological and syntactic disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 208–217, 2007.

[16] Caio Corro and Ivan Titov. Differentiable perturb-and-parse: Semi-supervised parsing with a structured variational autoencoder. *arXiv preprint arXiv:1807.09875*, 2018.

[17] Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Structured prediction theory based on factor graph complexity. *Advances in Neural Information Processing Systems*, 29, 2016.

[18] David Roxbee Cox and David Victor Hinkley. *Theoretical statistics*. CRC Press, 1979.

[19] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[21] Artyom Gadetsky, Kirill Struminsky, Christopher Robinson, Novi Quadrianto, and Dmitry Vetrov. Low-variance black-box gradient estimates for the plackett-luce distribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10126–10135, 2020.

[22] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR, 2017.

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[24] Will Grathwohl, Dami Choi, Yuhuai Wu, Geoff Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations*, 2018.

[25] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[26] Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. Stochastic optimization of sorting networks via continuous relaxations. In *International Conference on Learning Representations*, 2018.

[27] Jiatao Gu, Qi Liu, and Kyunghyun Cho. Insertion-based decoding with automatically inferred generation order. *Transactions of the Association for Computational Linguistics*, 7:661–676, 2019.

[28] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[29] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR 2017)*. OpenReview. net, 2017.

[30] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[31] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[32] Yoon Kim, Sam Wiseman, and Alexander M Rush. A tutorial on deep latent variable models of natural language. *arXiv preprint arXiv:1812.06834*, 2018.

[33] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

[34] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[35] Yi Lin. A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1):73–82, 2004.

[36] Ben London, Bert Huang, and Lise Getoor. Stability and generalization in structured prediction. *The Journal of Machine Learning Research*, 17(1):7808–7859, 2016.

[37] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[38] C Maddison, A Mnih, and Y Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of the international conference on learning Representations*. International Conference on Learning Representations, 2017.

[39] André FT Martins, Tsvetomila Mihaylova, Nikita Nangia, and Vlad Niculae. Latent structure models for natural language processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, 2019.

[40] Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *International Conference on Learning Representations*, 2018.

[41] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[42] Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. A general theory for structured prediction with smooth convex surrogates. *arXiv preprint arXiv:1902.01958*, 2019.

[43] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[44] Anton Osokin, Francis R Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *NIPS*, 2017.

[45] Max Benedikt Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, and Chris J Maddison. Gradient estimation with stochastic softmax tricks. In *NeurIPS 2020*, 2020.

[46] Judea Pearl. *Causality*. Cambridge university press, 2009.

[47] Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18:1–35, 2017.

[48] Lawrence Rabiner and Biinghwang Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[49] Harish G Ramaswamy, Shivani Agarwal, and Ambuj Tewari. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *NIPS*, 2013.

[50] D Raj Reddy et al. Speech understanding systems: A summary of results of the five-year research effort. *Department of Computer Science. Camegie-Mell University, Pittsburgh, PA*, 17:138, 1977.

[51] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?"explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[52] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[53] BTCGD Roller, C Taskar, and D Guestrin. Max-margin markov networks. *Advances in neural information processing systems*, 16:25, 2004.

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[55] Yasubumi Sakakibara, Michael Brown, Richard Hughey, I Saira Mian, Kimmen Sjölander, Rebecca C Underwood, and David Haussler. Stochastic context-free grammers for trna modeling. *Nucleic acids research*, 22(23):5112–5120, 1994.

[56] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):1–21, 2021.

[57] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[58] Noah A Smith. Linguistic structure prediction. *Synthesis lectures on human language technologies*, 4(2):1–274, 2011.

[59] K Struminskiy, A Klenitskiy, A Reshytko, D Egorov, A Shchepetnov, A Sabirov, D Vetrov, A Semenikhin, O Osmonalieva, and B Belozerov. Well log data standardization, imputation and anomaly detection using hidden markov models. In *Petroleum Geostatistics 2019*, volume 2019, pages 1–5. European Association of Geoscientists & Engineers, 2019.

[60] Kirill Struminsky, Stanislav Kruglik, Dmitry Vetrov, and Ivan Oseledets. A new approach for sparse bayesian channel estimation in scma uplink systems. In *2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)*, pages 1–5. IEEE, 2016.

[61] Kirill Struminsky, Simon Lacoste-Julien, and Anton Osokin. Quantifying learning guarantees for convex but inconsistent surrogates. In *NeurIPS*, 2018.

[62] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[63] Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin markov networks. *Advances in neural information processing systems*, 16, 2003.

[64] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

[65] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, Yasemin Altun, and Yoram Singer. Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6(9), 2005.

[66] George Tucker, Andriy Mnih, Chris J Maddison, Dieterich Lawson, and Jascha Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. *arXiv preprint arXiv:1703.07370*, 2017.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[68] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[69] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

[70] Yufeng Wang, Shidong Zhou, Limin Xiao, Xiujun Zhang, and Jin Lian. Sparse bayesian learning based user detection and channel estimation for scma uplink systems. In *2015 International Conference on Wireless Communications & Signal Processing (WCSP)*, pages 1–5. IEEE, 2015.

[71] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[72] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

[73] John I Yellott Jr. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144, 1977.

[74] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NIPS*, 2014.