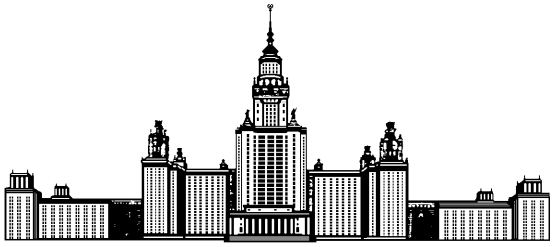

Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

ДИПЛОМНАЯ РАБОТА

Вычислительные методы приближенного подсчета нормировочной константы марковского случайного поля

Выполнил:
студент 517 группы
Новиков Александр Витальевич

Научный руководитель:
к.ф-м.н., доцент
Ветров Дмитрий Петрович

Москва, 2015

Содержание

Аннотация

В данной работе предлагаются три метода приближенного подсчёта нормировочной константы марковского случайного поля. Первый метод основан на разложении Tensor Train, которое позволяет компактно хранить тензор и эффективно применять к нему операции линейной алгебры. Свойства этого разложения используются для оценки нормировочной константы и получения теоретических гарантий на точность работы метода. Второй метод ускоряет и обобщает недавно предложенный метод n-окрестностей до применимости к произвольным марковским случайным полям. Третий метод основан на разложении показательной функции в определении нормировочной константы в ряд Тейлора. Предложенные методы сравниваются друг с другом и с доступными аналогами на моделях Изинга и на задаче обучения ограниченной машины Больцмана по выборке MNIST.

1 Введение

Вероятностные графические модели являются удобным инструментом для решения задач в таких областях, как компьютерное зрение, машинное обучение, анализ социальных сетей и т. д. Одно из их основных достоинств — это возможность компактно задать ненормированное распределение на сотни переменных в виде произведения факторов низких порядков. Это позволяет, с одной стороны, учесть сложные зависимости в данных, а с другой стороны точно или приближенно вычислять статистики полученного распределения.

Многие возникающие задачи становятся сложными, если фактор-граф содержит циклы. В работе рассматриваются ненаправленные графические модели (марковские случайные поля, MRF), и поднимается актуальная для них задача оценки нормировочной константы распределения по параметрам модели. Решение этой задачи особенно важно при настройке параметров графической модели по тренировочной выборке (обучение по методу максимального правдоподобия [24]). В последние годы для этой задачи предложено множество различных методов, но, тем не менее, её пока рано называть полностью решенной.

Совместное распределение дискретных переменных можно задать через многомерный массив (тензор) ненормированной вероятности. Тогда подсчет нормировочной константы сводится к суммированию всех элементов тензора. Для хранения и обработки тензора в явном виде требуется экспоненциальное количество памяти и вычислительных ресурсов. Тензорные разложения позволяют компактно хранить тензор и эффективно совершать над ним различные операции.

В дипломной работе рассматривается недавно предложенное И. Оселедцом [18] разложение тензорного поезда (TT-разложение или TT-формат). Алгоритмы TT-разложения хранят тензор в специальном формате, который позволяет эффективно оперировать над тензорами. Эффективность TT-разложения определяется величиной TT-рангов тензора. В работе показано, что тензор энергии (минус логарифма вероятности) марковского случайного поля обладает низкими TT-рангами, а тензор вероятности — высокими TT-рангами. Для подсчета нормировочной константы в условиях высоких TT-рангов тензора вероятности предложен метод, основанный на факторизации графической модели. Факторы графической модели переводятся

в TT-формат и комбинируются таким образом, чтобы оценить нормировочную константу, не строя сам тензор вероятности.

Так же в работе рассматривается предложенный в 2014 году метод n-окрестностей [10], придуманный для оценки нормировочной константы (или, как её называют в физической литературе, «статистической суммы») модели Изинга[1]. Основная идея метода n-окрестностей – разбить множество значений энергии на подмножества и приблизить эмпирическую функцию распределения энергий в каждом подмножестве с помощью нормального распределения. В дипломной работе предлагается отказаться от рассмотрения отдельных подмножеств и рассматривать множество всех энергий целиком. Полученные формулы позволяют применять такой метод к произвольному MRF (тогда как метод n-окрестностей предложен только для модели Изинга), а использование полного множества энергий в ряде случаев позволяет получить более точные результаты.

Формулы, выведенные для обобщения метода n-окрестностей, используются в третьем предложенном методе, основанном на разложении показательной функции в определении нормировочной константы в ряд Тейлора.

Основной вклад настоящей работы заключается в следующем:

- Предложен алгоритм оценки нормировочной константы через TT-разложение факторов графической модели, который не строит TT-представление тензора ненормированного совместного распределения.
- Приведены теоретические гарантии на точность оценки нормировочной константы.
- Предложено обобщение метода n-окрестностей, применимое к произвольным марковским случайным полям и обладающее более высокой скоростью работы.
- Предложен простой и быстрый метод оценки нормировочной константы через разложение в ряд Тейлора.

[1]Модель Изинга – графическая модель, используемая в статистической физики для моделирования магнитных свойств вещества.

Настоящая дипломная работа организована следующим образом. В разделе 4 приведён краткий обзор тензорного разложения Tensor Train, в разделе 5 поясняется его связь с MRF, а в разделе 6 описывается предлагаемый метод оценки нормировочной константы, основанный на разложении Tensor Train. В разделе 7 приведён краткий обзор недавно предложенного в физической литературе метода для подсчета нормировочной константы модели Изинга, а в разделе 8 описывается предлагаемое обобщение и ускорение данного метода, которое называется методом «Суммировать всё». В разделе 9 предлагается простой и быстрый метод для оценки нормировочной константы, который основан на разложении показательной функции в определении ненормированного распределения Гиббса в ряд Тейлора. Раздел 10 содержит результаты экспериментального сравнения предлагаемых подходов друг с другом и с лучшими аналогами.

2 Обзор литературы

Обзор можно разделить на три основных направления: тензорные разложения, позволяющие компактно представить тензор, использование тензорных методов в графических моделях и различные алгоритмы оценки нормировочной константы MRF.

В отличие от классических тензорных форматов (канонический формат [1] и формат Таккера [22]), для TT-разложения существуют устойчивые алгоритмы построения, а само TT-разложение не подвержено проклятию размерности. Альтернативой ему является иерархический формат Таккера [5, 3], который также позволяет устойчиво искать компактное представление тензора. TT-формат эквивалентен иерархическому формату Таккера с линейным деревом размерностей, что является преимуществом с алгоритмической точки зрения: большинство алгоритмов TT-формата значительно проще аналогов для иерархического формата Таккера.

В статьях по графическим моделям тензорные подходы часто упоминаются применительно к восстановлению структуры модели. Джернит и др.; Мария Иштева и др. принимают локальные решения, основываясь на свойствах четырехмерных тен-

зоров [8, 7]. Сонг и др. используют иерархические тензорные разложения для восстановления параметров графической модели [21].

Можно выделить следующие группы методов подсчета нормировочной константы: методы, основанные на генерации выборки (например Annealed Importance Sampling [15, 4]), методы передачи сообщений (Loopy Belief Propagation [11] и его многочисленные модификации), методы, основанные на минимизации KL-дивергенции (Mean Field [24] и Expectation Propagation [13]), методы декомпозиции графов (Tree-Reweighted Message-passing [23]) и методы, основанные на минимизации энергии (randomized MAP-predictors [6]).

3 Обозначения и формальная постановка задачи

В данной статье часто будет использоваться понятие многомерного массива вещественных чисел. Одномерные массивы будем называть *векторами*, двумерные — *матрицами*, а массивы большей размерности будем называть *тензорами*. Для обозначения векторов будут использоваться маленькие жирные буквы ($\boldsymbol{a}$), а для матриц и тензоров большие жирные буквы ($\boldsymbol{A}$).

Будем рассматривать все массивы как функции их индексов: $a(i) = a_i$, $A(x_1, x_2)$, $A(\boldsymbol{x}) = A(x_1, \ldots, x_n)$, где n – это размерность тензора $\mathbf{A}$.

Символом $\|\cdot\|_F$ будем обозначать норму Фробениуса матриц и её обобщение на случай тензоров:

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{x_1,\ldots,x_n} A^2(x_1, \ldots, x_n)}.$$

Под $\|\cdot\|_2$ будем понимать евклидову норму векторов и спектральную норму матриц: $\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$.

В работе будут использоваться несколько различных матричных произведений: произведение Адамара (символ "$\odot$"), произведение Кронекера (символ "$\otimes$") и обычное матричное произведение (символ "$\cdot$").

Введём теперь формально определение марковского случайного поля. Рассмотрим гиперграф $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ с конечным множеством вершин $\mathcal{V}$ и гиперребер $\mathcal{E}$. Пусть все вершины пронумерованы от 1 до n, а все гиперребра от 1 до m.

Сопоставим каждой вершине $i = 1, \ldots, n$ переменную x_i принимающую значения из множества $\mathcal{X}_i$ мощности d_i: $|\mathcal{X}_i| = d_i$ (например $\mathcal{X}_i = \{1, \ldots, d_i\}$). Обозначим символом $\mathcal{X}$ совместную область определения переменных $x_1, \ldots, x_n$: $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$. Обозначим множество вершин, входящих в гиперребро $\ell = 1, \ldots, m$, через $\boldsymbol{x}^\ell$. Каждому гиперребру $\ell = 1, \ldots, m$ сопоставим вещественнозначную функцию $\boldsymbol{\Theta}_\ell$, определенную на совместной области определения переменных из $\boldsymbol{x}^\ell$. Функцию $\boldsymbol{\Theta}_\ell$ будем называть *потенциалом*.

Функцией энергии марковского случайного поля (Markov random field, MRF) заданного на гиперграфе $\mathcal{G}$, назовем сумму всех потенциалов: $E(\boldsymbol{x}) = \sum_{\ell=1}^{m} \Theta_\ell(\boldsymbol{x}^\ell)$.

Экспонента от минус энергии задает *ненормированное распределение Гиббса*: $\widehat{P}(\boldsymbol{x}) = \exp(-E(\boldsymbol{x}))$. Для обозначения нормировочной константы будет использоваться символ Z:

$$Z = \sum_{\boldsymbol{x}} \widehat{P}(\boldsymbol{x}). \tag{1}$$

В дипломной работе рассматривается следующая задача: по заданному MRF (т. е. гиперграфу и функции энергии) требуется оценить величину логарифма нормировочной константы $\ln(Z)$.

4 TT-разложение

Будем говорить, что n-мерный тензор $\mathbf{A}$ представлен в TT-формате, если для всех размерностей $i = 1, \ldots, n$ и всех значений индексов по этой размерности $x_i = 1, \ldots, d_i$ ($d = \max_{i=1,\ldots,n} d_i$) существуют матрицы $G_i^{\mathbf{A}}[x_i]$, такие, что каждый элемент тензора $\mathbf{A}$ представим в виде произведения матриц:

$$A(x_1, \ldots, x_n) = \boldsymbol{G}_1^{\mathbf{A}}[x_1]\boldsymbol{G}_2^{\mathbf{A}}[x_2] \ldots \boldsymbol{G}_n^{\mathbf{A}}[x_n]. \tag{2}$$

При этом все матрицы относящиеся к одной и той же размерности i должны иметь одинаковые размеры $\mathrm{r}_{i-1}(\mathbf{A}) \times \mathrm{r}_i(\mathbf{A})$. Чтобы результат матричного произведения (2) был равен числу, положим $\mathrm{r}_0(\mathbf{A}) = \mathrm{r}_n(\mathbf{A}) = 1$. Последовательность $\{\mathrm{r}_i(\mathbf{A})\}_{i=0}^n$ будем называть *TT-рангами* тензора $\mathbf{A}$, а максимальный элемент последовательности — *максимальным TT-рангом* тензора $\mathbf{A}$: $\mathrm{r}(\mathbf{A}) = \max_{i=0,\ldots,n} \mathrm{r}_i(\mathbf{A})$. Набор матриц $\boldsymbol{G}_i^{\mathbf{A}}$, со-

ответствующих одному измерению, называется *TT-ядром* тензора $\mathbf{A}$. Представление тензора в TT-формате будем называть *TT-разложением* или *TT-представлением*.

Для любого n-мерного тензора $\mathbf{A}$ существует TT-представление с максимальным TT-рангом $\mathrm{r}(\mathbf{A}) \leq d^{\frac{n}{2}}$ (см. теорему 2.1 Оселедца [18]). Отметим, что TT-представление тензора не единственно.

Для обозначения (α_{i-1}, α_i)-ого элемента матрицы $\boldsymbol{G}_i^{\mathbf{A}}[x_i]$ будет использоваться символ $G_i^{\mathbf{A}}[x_i](\alpha_{i-1}, \alpha_i)$. Пользуясь определением произведения матриц, можно переписать формулу (2) через элементы TT-ядер:

$$A(\boldsymbol{x}) = \sum_{\alpha_0,\ldots,\alpha_n} G_1^{\mathbf{A}}[x_1](\alpha_0, \alpha_1) \ldots G_n^{\mathbf{A}}[x_n](\alpha_{n-1}, \alpha_n). \tag{3}$$

Для хранения всех элементов тензора $\mathbf{A}$ требуется $\prod_{i=1}^{n} d_i$ ячеек памяти, тогда как хранение $\mathbf{A}$ в TT-формате требует $\sum_{i=1}^{n} d_i \, \mathrm{r}_{i-1}(\mathbf{A}) \, \mathrm{r}_i(\mathbf{A})$. Таким образом, TT-представление тензора с низкими TT-рангами существенно компактнее перечисления его элементов.

Существуют две различные постановки задачи перевода тензора в TT-формат: точное TT-представление (алгоритм TT-SVD [18], применимый для небольших тензоров), и построение приближенного TT-представления по небольшому подмножеству элементов тензора. Наилучшим из алгоритмов второго класса в настоящее время является метод AMEn-cross [2].

Одним из основных достоинств TT-формата является возможность эффективно применять различные операции к тензорам в TT-формате: умножение тензора на константу, добавление константы к тензору, поточечное сложение и умножение тензоров (результат этих операций — это тензор в TT-формате с возросшими TT-рангами); подсчет глобальных характеристик тензора, таких как сумма всех элементов или норма Фробениуса. Обзор операций, которые используются в этой работе, приведен в таблице 1 (детальный обзор проведен Оселедцом [18]).

Применение операций к TT-тензорам увеличивает TT-ранги даже в том случае, когда существует низкоранговое TT-представление результата. Чтобы контролировать рост TT-рангов существует операция TT-округления. По TT-представлению тензора $\mathbf{A}$ и относительной точности $\varepsilon \geq 0$ операция TT-округления $\mathrm{round}(\mathbf{A}, \varepsilon)$ найдет тензор $\widehat{\mathbf{A}}$ в TT-формате который, во-первых, достаточно близок к тензору $\mathbf{A}$:

Таблица 1: Операции, которые можно эффективно выполнять над тензорами в TT-формате. Для каждой операции указана её вычислительная сложность и TT-ранг результата для ситуаций, когда результат является тензором в TT-формате.

Операция	Ранг результата	Вычислительная сложность
$\mathbf{C} = \mathbf{A} \cdot \text{const}$	$\mathrm{r}(\mathbf{C}) = \mathrm{r}(\mathbf{A})$	$O(d\,\mathrm{r}(\mathbf{A}))$
$\mathbf{C} = \mathbf{A} + \text{const}$	$\mathrm{r}(\mathbf{C}) = \mathrm{r}(\mathbf{A}){+}1$	$O(nd\,\mathrm{r}^2(\mathbf{A}))$
$\mathbf{C} = \mathbf{A} + \mathbf{B}$	$\mathrm{r}(\mathbf{C}) \le \mathrm{r}(\mathbf{A}){+}\mathrm{r}(\mathbf{B})$	$O(nd\,(\mathrm{r}(\mathbf{A}) + \mathrm{r}(\mathbf{B}))^2)$
$\mathbf{C} = \mathbf{A} \odot \mathbf{B}$	$\mathrm{r}(\mathbf{C}) \le \mathrm{r}(\mathbf{A})\,\mathrm{r}(\mathbf{B})$	$O(nd\,\mathrm{r}^2(\mathbf{A})\,\mathrm{r}^2(\mathbf{B}))$
$\boldsymbol{c} = \boldsymbol{M}\boldsymbol{b}$	$\mathrm{r}(\boldsymbol{c}) \le \mathrm{r}(M)\,\mathrm{r}(\boldsymbol{b})$	$O(nd^2\,\mathrm{r}^2(M)\,\mathrm{r}^2(\boldsymbol{b}))$
$\operatorname{sum} \mathbf{A}$	–	$O(nd\,\mathrm{r}^2(\mathbf{A}))$
$\|\mathbf{A}\|_F$	–	$O(nd\,\mathrm{r}^3(\mathbf{A}))$
$\mathbf{C} = \operatorname{round}(\mathbf{A}, \varepsilon)$	$\mathrm{r}(\mathbf{C}) \le \mathrm{r}(\mathbf{A})$	$O(nd\,\mathrm{r}^3(\mathbf{A}))$

$\|\mathbf{A} - \widehat{\mathbf{A}}\|_F \le \varepsilon \|\mathbf{A}\|_F$ и, во-вторых, обладает минимальными TT-рангами среди всех тензоров $\mathbf{B}$: $\|\mathbf{A} - \mathbf{B}\|_F \le \frac{\varepsilon}{\sqrt{n-1}} \|\mathbf{A}\|_F$. Наличие операции TT-округления позволяет применять последовательность операций к тензорам (например, округляя результат после применения каждой операции), контролируя рост TT-рангов.

Для повышения эффективности работы с векторами и матрицами специальным образом вводятся понятия TT-формата вектора и TT-формата матрицы. Пусть существует отображение между индексами вектора $\boldsymbol{b} \in \mathbb{R}^q$ и n-мерными векторами $\boldsymbol{y} = (y_1, \ldots, y_n)$[2]. *TT-представлением вектора* $\boldsymbol{b}$ называется TT-представление тензора $\mathbf{B}(y_1, \ldots, y_n)$, содержащего все элементы $\boldsymbol{b}$.

Рассмотрим пример. Пусть вектор $\boldsymbol{b}$ содержит 18 элементов. Его можно интерпретировать как (например) трехмерный тензор $\mathbf{B}$:

$$
\begin{aligned}
&1 \leftrightarrow (1,1,1) && B(1,1,1) = b(1),\\
&2 \leftrightarrow (1,1,2) && B(1,1,2) = b(2),\\
&\ldots && \ldots\\
&18 \leftrightarrow (2,3,3) && B(2,3,3) = b(18).
\end{aligned}
$$

[2] Количество элементов в векторе $\boldsymbol{b}$ равно $q = \prod_{i=1}^{n} d_i$.

Здесь составной индекс $\boldsymbol{y} = (y_1, y_2, y_3)$ нумерует элементы вектора $\boldsymbol{b}$.

Определим понятие TT-формата для матриц. Пусть существует отображение между индексами строк и столбцов матрицы $\boldsymbol{M}$ в n-мерные вектора $\boldsymbol{x}$ и $\boldsymbol{y}$ соответственно. Переупорядочим размерности и представим получившийся тензор в TT-формате:

$$M((x_1, y_1), \ldots, (x_n, y_n)) = \boldsymbol{G}_1^{\boldsymbol{M}}[x_1, y_1] \ldots \boldsymbol{G}_n^{\boldsymbol{M}}[x_n, y_n],$$

где $\boldsymbol{G}_i^{\mathbf{M}}$, $i = 1, \ldots, n$ — это TT-ядра, а $\boldsymbol{G}_i^{\mathbf{M}}[x_i, y_i]$ — матрицы. *TT-представлением матрицы* $\boldsymbol{M}$ будем называть TT-представление тензора $\boldsymbol{M}$. Отметим, что матрица в TT-формате не обязана быть квадратной, т. к. x_i и y_i могут принимать разное число возможных значений

Для матрицы $\boldsymbol{M}$ и вектора $\boldsymbol{b}$ представленных в TT-формате можно эффективно вычислять произведение $\boldsymbol{c} = \boldsymbol{M}\boldsymbol{b}$ (если соответствующие размерности совпадают). Результатом этой операции является вектор $\boldsymbol{c}$ в TT-формате с рангами равными произведению рангов $\boldsymbol{M}$ и $\boldsymbol{b}$: $\mathrm{r}_i(\boldsymbol{c}) = \mathrm{r}_i(\boldsymbol{M})\,\mathrm{r}_i(\boldsymbol{b})$.

Наличие специального определения TT-формата для векторов и матриц позволяет применять операции линейной алгебры к задачам большого размера. Например, для поиска минимального элемента тензора можно вытянуть его в диагональную матрицу и применить приближенный метод поиска минимальных собственных значений, основанный на алгоритме DMRG [9].

5 Тензоры марковского случайного поля

Как энергия, так и ненормированная вероятность являются n-мерными тензорами. Потенциалы и факторы можно интерпретировать как n-мерные тензоры, если добавить в них переменные, от которых они зависят несущественно: $\Theta_\ell(\boldsymbol{x}^\ell) = \Theta_\ell(\boldsymbol{x})$. Тензоры энергии и вероятности можно определить следующим образом: $\mathbf{E} = \sum_{\ell=1}^{m} \boldsymbol{\Theta}_\ell$ и $\widehat{\mathbf{P}} = \bigodot_{\ell=1}^{m} \boldsymbol{\Psi}_\ell$.

Традиционно для компактного представления энергии используется набор потенциалов MRF. TT-разложение является альтернативным способом компактного представления тензора энергии. Эффективность применения TT-формата для тензора энергии и тензора вероятности обсуждается в разделах 5.1 и 5.2.

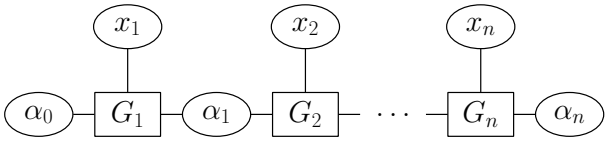

Рис. 1: Графическая модель, полученная в результате TT-разложения тензора вероятности $P(x_1, \ldots, x_n)$.

TT-представление тензора совместного распределения $\mathbf{P} = \widehat{\mathbf{P}}/Z$ обладает особой интерпретацией. Переменные α_i, $i = 0, \ldots, n$, (см. (3)) можно интерпретировать как скрытые переменные в графической модели типа цепочка (рис. 1). Маргинализация по ним дает исходную вероятность:

$$P(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}} P(\boldsymbol{x}, \boldsymbol{\alpha}), \tag{4}$$

где $P(\boldsymbol{x}, \boldsymbol{\alpha})$ — совместное распределение вероятностей на векторы $\boldsymbol{x}$ и $\boldsymbol{\alpha}$. Каждая добавленная переменная α_i может принимать значения от 1 до соответствующего TT-ранга $\mathrm{r}_i(\mathbf{P})$.

5.1 TT-формат для тензора энергии

В этом разделе предложен алгоритм для представление тензора энергии $\mathbf{E}$ в TT-формате. Предлагаемый алгоритм существенно превосходит AMEn-cross по скорости и точности.

Алгоритм состоит из трех основных шагов.

Шаг 1. Найдем TT-представление для каждого потенциала $\Theta_\ell(\boldsymbol{x}^\ell)$, который будем интерпретировать как тензор зависящий только от переменных $\boldsymbol{x}^\ell$. Обычно потенциалы существенно зависят лишь от небольшого числа переменных и можно искать их TT-представление с помощью алгоритма TT-SVD.

Шаг 2. Добавим в каждый тензор $\boldsymbol{\Theta}_\ell$ несущественные переменные $\boldsymbol{x} \setminus \boldsymbol{x}^\ell$, сделав его n-мерным тензором.

Пусть тензор $\boldsymbol{\Theta}_\ell$ существенно зависит от p переменных: $\boldsymbol{x}^\ell = (x_{i_1}, \ldots, x_{i_p})$, $i_1 < i_2 < \ldots < i_p$. После первого шага алгоритма получим следующее TT-пред-

ставление $\Theta_\ell(\boldsymbol{x}^\ell)$:

$$\Theta_\ell(x_{i_1}, \ldots, x_{i_p}) = \overline{\boldsymbol{G}}_1^{\boldsymbol{\Theta}_\ell}[x_{i_1}] \ldots \overline{\boldsymbol{G}}_p^{\boldsymbol{\Theta}_\ell}[x_{i_p}], \tag{5}$$

где $\overline{\boldsymbol{G}}_k^{\boldsymbol{\Theta}_\ell}[x_{i_k}]$ — матрица размера $\bar{\mathrm{r}}_{k-1} \times \bar{\mathrm{r}}_k$. Чтобы добавить несущественные размерности, можно, например, положить недостающие ядра равными набору единичных матриц подходящего размера[3]:

$$\begin{aligned}
\boldsymbol{G}_1^{\boldsymbol{\Theta}_\ell}[x_1] &\equiv \ldots \equiv \boldsymbol{G}_{i_1-1}^{\boldsymbol{\Theta}_\ell}[x_{i_1-1}] \equiv \boldsymbol{I}_{\bar{\mathrm{r}}_0} = \boldsymbol{I}_1, \\
\boldsymbol{G}_{i_1+1}^{\boldsymbol{\Theta}_\ell}[x_{i_1+1}] &\equiv \ldots \equiv \boldsymbol{G}_{i_2-1}^{\boldsymbol{\Theta}_\ell}[x_{i_2-1}] \equiv \boldsymbol{I}_{\bar{\mathrm{r}}_1}, \\
&\ldots \\
\boldsymbol{G}_{i_p+1}^{\boldsymbol{\Theta}_\ell}[x_{i_p+1}] &\equiv \ldots \equiv \boldsymbol{G}_n^{\boldsymbol{\Theta}_\ell}[x_n] \equiv \boldsymbol{I}_{\bar{\mathrm{r}}_p} = \boldsymbol{I}_1,
\end{aligned}$$

где $\boldsymbol{G}_{i_k}^{\boldsymbol{\Theta}_\ell}[x_{i_k}] = \overline{\boldsymbol{G}}_k^{\boldsymbol{\Theta}_\ell}[x_{i_k}]$. Полученные таким образом ТТ-ядра задают ТТ-формат тензора $\Theta_\ell(\boldsymbol{x})$. Отметим, что в процессе данного преобразования максимальный ТТ-ранг $\boldsymbol{\Theta}_\ell$ не изменяется.

Шаг 3. Просуммируем полученные на шаге 2 ТТ-тензоры $\boldsymbol{\Theta}_\ell$, чтобы получить тензор энергии:

$$\mathbf{E} = \sum_{\ell=1}^{m} \boldsymbol{\Theta}_\ell. \tag{6}$$

Теорема 1 содержит верхнюю оценку на максимальный ТТ-ранг тензора, построенного описанным алгоритмом.

Теорема 1. *Если порядок каждого потенциала не превосходит p, то предложенный алгоритм строит ТТ-представление тензора энергии $\mathbf{E}$ с максимальным ТТ-рангом, ограниченным сверху следующим образом:*

$$\mathrm{r}(\mathbf{E}) \leq d^{\frac{p}{2}} \cdot m. \tag{7}$$

Теорема 1 показывает, что можно построить точное низкоранговое ТТ-представления тензора энергии. Отметим, что после шага 3 к тензору энергии можно применить процедуру ТТ-округления, чтобы уменьшить ТТ-ранги, если это возможно.

[3] $\boldsymbol{I}_k$ используется для обозначения единичной матрицы размера $k \times k$.

Доказательство теоремы 1. Оценим, какими TT-рангами обладают факторы после шага 1 алгоритма. Из теоремы 2.1 [18] следует, что TT-ранги тензора не превосходят $d^{\frac{n}{2}}$, где n — это размерность тензора, а d — максимальное число возможных значений, которые может принимать каждая переменная x_i. По условию теоремы, порядок каждого потенциала не превосходит p, а значит после шага 1 верна следующая оценка:

$$\mathrm{r}(\mathbf{\Theta}_\ell) \leq d^{\frac{p}{2}}. \tag{8}$$

На шаге 2 максимальный TT-ранг потенциалов не возрастает, а значит после шага 2 неравенство (8) всё ещё выполнено.

При суммировании тензоров на шаге 3, TT-ранги растут аддитивно. Чтобы завершить доказательство, осталось вспомнить, что всего есть m потенциалов. □

5.2 TT-формат для тензора вероятности

Алгоритм из раздела 5.1 легко адаптировать для построения TT-представления тензора ненормированной вероятности $\widehat{\mathbf{P}}$.

На шагах 1 и 2 алгоритма вместо потенциалов $\mathbf{\Theta}_\ell$ следует работать с факторами $\mathbf{\Psi}_\ell$. На шаге 3 вместо суммы нужно вычислить поэлементное произведение:

$$\widehat{\mathbf{P}} = \bigodot_{\ell=1}^{m} \mathbf{\Psi}_\ell. \tag{9}$$

Данный алгоритм построит точное TT-представление тензора $\widehat{\mathbf{P}}$. Тем не менее, TT-ранги $\widehat{\mathbf{P}}$ экспоненциально зависят от количества вершин (рис. 2), что делает TT-представление тензора вероятности неприменимым на больших задачах. На рис. 2 изображены ранги точного TT-представления тензора энергии и вероятности для MRF разного размера. TT-ранги тензора вероятности остаются экспоненциально большими и после округления $\widehat{\mathbf{P}}$ с точностью $\varepsilon = 10^{-8}$. Таким образом, точного низкорангового TT-представления $\widehat{\mathbf{P}}$ не существует.

6 Тензорный метод

Естественный подход к подсчету нормировочной константы состоит в представлении всего тензора ненормированной вероятности $\widehat{\mathbf{P}}$ в TT-формате и в последующем

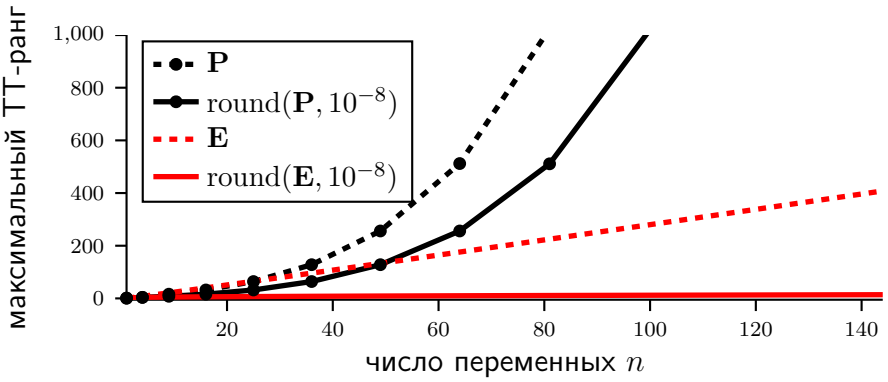

Рис. 2: Максимальный TT-ранг тензора энергии $\mathbf{E}$ и тензора ненормированной вероятности $\widehat{\mathbf{P}}$ для модели Изинга с температурой равной 10 и весами парных потенциалов равными 1. Детали см. в разделе 10.2.

суммировании всех его элементов. На практике тензор $\widehat{\mathbf{P}}$ обладает экспоненциально большими TT-рангами, и работа с его TT-представлением становиться неэффективной. С другой стороны, каждый отдельный фактор графической модели можно точно представить в TT-формате. В этом разделе предложен алгоритм подсчета нормировочный константы, который работает с TT-представлением отдельных факторов $\mathbf{\Psi}_\ell$, не строя TT-представление всего тензора $\widehat{\mathbf{P}}$.

6.1 Алгоритм

Пусть все факторы графической модели уже представлены в TT-формате (см. раздел 5.2):

$$\Psi_\ell(\boldsymbol{x}^\ell) = \Psi_\ell(\boldsymbol{x}) = \boldsymbol{G}_1^{\mathbf{\Psi}_\ell}[x_1] \dots \boldsymbol{G}_n^{\mathbf{\Psi}_\ell}[x_n]. \tag{10}$$

Далее символ $G_i^\ell[x_i](\alpha_{i-1}^\ell, \alpha_i^\ell)$ будет использоваться как сокращенное обозначение для $G_i^{\mathbf{\Psi}_\ell}[x_i](\alpha_{i-1}^{\mathbf{\Psi}_\ell}, \alpha_i^{\mathbf{\Psi}_\ell})$.

По определению, нормировочная константа Z вычисляется как сумма значений ненормированного распределения на всех конфигурациях:

$$Z = \sum_{\boldsymbol{x}} \prod_{\ell=1}^{m} \underbrace{\Psi_\ell(\boldsymbol{x})}_{\in \mathbb{R}} = \sum_{x_1,\dots,x_n} \bigotimes_{\ell=1}^{m} \left(\boldsymbol{G}_1^\ell[x_1] \dots \boldsymbol{G}_n^\ell[x_n]\right).$$

Второе равенство выполнено, т. к. кронекерово произведение чисел (матриц размера 1×1) эквивалентно обычному произведению.

Пользуясь свойством смешанного произведения $\boldsymbol{AC} \otimes \boldsymbol{BD} = (\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D})$, преобразуем выражение для Z:

$$Z = \sum_{x_1,\ldots,x_n} \left(\boldsymbol{G}_1^1[x_1] \otimes \ldots \otimes \boldsymbol{G}_1^m[x_1]\right) \ldots \left(\boldsymbol{G}_n^1[x_n] \otimes \ldots \otimes \boldsymbol{G}_n^m[x_n]\right).$$

Обозначим кронекерово произведение матриц $\boldsymbol{G}_i^\ell[x_i]$ через $\boldsymbol{A}_i[x_i]$:

$$\boldsymbol{A}_i[x_i] = \boldsymbol{G}_i^1[x_i] \otimes \ldots \otimes \boldsymbol{G}_i^m[x_i].$$

Для любого значения x_i матрица $\boldsymbol{A}_i[x_i]$ имеет размеры $(\mathrm{r}_{i-1}(\boldsymbol{\Psi}_1)\ldots\mathrm{r}_{i-1}(\boldsymbol{\Psi}_m)) \times (\mathrm{r}_i(\boldsymbol{\Psi}_1)\ldots\mathrm{r}_i(\boldsymbol{\Psi}_m))$. Значение её элементов выражается через элементы матриц $\boldsymbol{G}_i^\ell[x_i]$:

$$A_i[x_i](\alpha_{i-1}^1,\ldots,\alpha_{i-1}^m;\alpha_i^1,\ldots,\alpha_i^m) = G_i^1[x_i](\alpha_{i-1}^1,\alpha_i^1)\ldots G_i^m[x_i](\alpha_{i-1}^m,\alpha_i^m).$$

Таким образом, матрица $\boldsymbol{A}_i[x_i]$ представлена в TT-формате, а её TT-ранг равен одному (т к. $G_i^\ell[x_i](\alpha_{i-1}^\ell,\alpha_i^\ell)$ — это матрица размера 1×1).

Представим нормировочную константу Z в виде произведения n матриц:

$$Z = \sum_{x_1,\ldots,x_n} \boldsymbol{A}_1[x_1]\ldots\boldsymbol{A}_n[x_n] = \left(\sum_{x_1} \boldsymbol{A}_1[x_1]\right)\ldots\left(\sum_{x_n} \boldsymbol{A}_n[x_n]\right) = \boldsymbol{B}_1\ldots\boldsymbol{B}_n,$$

где

$$\boldsymbol{B}_i = \sum_{x_i=1}^{d_i} \boldsymbol{A}_i[x_i].$$

TT-преставление матрицы $\boldsymbol{B}_i$ можно получить просуммировав d_i матриц в TT-формате. Все матрицы $\boldsymbol{B}_i$ можно построить и держать в оперативной памяти, т. к. TT-ранги $\boldsymbol{B}_i$ не превосходят d_i.

Матрицы $\boldsymbol{B}_1$ и $\boldsymbol{B}_n$ являются вектором-строкой и вектором-столбцом соответственно, а значит результат произведения $\boldsymbol{B}_1\ldots\boldsymbol{B}_n$ — это число. Построив TT-матрицы $\boldsymbol{B}_i$, их можно перемножать, округляя результат после каждого умножения (см. алгоритм 1). Параметр ε контролирует баланс между точностью и скоростью работы алгоритма.

Помимо нормировочной константы Z, предложенный метод также позволяет найти маргинальные распределения на переменные графической модели. Ненормированное маргинальное распределение $\hat{P}_i(x_i)$ вычисляется следующим образом: $\hat{P}_i(x_i) = \boldsymbol{B}_1\ldots\boldsymbol{B}_{i-1}\boldsymbol{A}_i[x_i]\boldsymbol{B}_{i+1}\ldots\boldsymbol{B}_n$. При этом произведения $\boldsymbol{B}_1\ldots\boldsymbol{B}_{i-1}$ и $\boldsymbol{B}_{i+1}\ldots\boldsymbol{B}_n$, $i =$

Алгоритм 1 Подсчет нормировочной константы Z

Вход: факторы $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_m$, точность округления ε

Выход: оценка нормировочной константы $\widehat{Z}$

для $\ell := 1$ **до** m
  Найти ТТ-ядра $\boldsymbol{G}_1^\ell, \ldots, \boldsymbol{G}_n^\ell$ для тензора $\boldsymbol{\Psi}_\ell$
Инициализировать $\boldsymbol{f}_{n+1} := 1$
для $i := n$ **до** 1
  Инициализировать $\boldsymbol{B}_i := 0$
  для $x_i := 1$ **до** d_i
    Построить ТТ-матрицу $\boldsymbol{A}_i[x_i] = \bigotimes_{\ell=1}^m \boldsymbol{G}_i^\ell[x_i]$
    $\boldsymbol{B}_i := \boldsymbol{B}_i + \boldsymbol{A}_i[x_i]$
  $\overline{\boldsymbol{f}_i} := \boldsymbol{B}_i \cdot \boldsymbol{f}_{i+1}$
  $\boldsymbol{f}_i := \mathrm{round}(\overline{\boldsymbol{f}_i}, \varepsilon)$
$\widehat{Z} := \boldsymbol{f}_1$

$1, \ldots, n$ можно предрассчитать за $2(n-1)$ умножений ТТ-матриц. Дополнительно рассчитав все произведения вида $\boldsymbol{B}_i \ldots \boldsymbol{B}_j$, $1 \le i < j \le n$, можно вычислять маргинальное распределение на любое подмножество переменных.

6.2 Анализ алгоритма 1

В этом разделе предоставлены теоретические гарантии точности оценки нормировочной константы алгоритмом 1.

Обозначим оценку произведения ТТ-матриц $\{\boldsymbol{B}_j\}_{j=i}^n$ за $\boldsymbol{f}_i$, $\boldsymbol{f}_n = \boldsymbol{B}_n$, $\widehat{Z} = \boldsymbol{f}_1$. Умножая ТТ-матрицу на ТТ-вектор и применяя ТТ-округление, получаем $\boldsymbol{f}_i$

$$\boldsymbol{f}_i = \mathrm{round}(\boldsymbol{B}_i \boldsymbol{f}_{i+1}, \varepsilon),$$

где точность ТТ-округления контролирует относительную точность по евклидовой норме[4]:

$$\|\boldsymbol{B}_i \boldsymbol{f}_{i+1} - \boldsymbol{f}_i\|_2 \le \varepsilon \|\boldsymbol{B}_i \boldsymbol{f}_{i+1}\|_2 . \tag{11}$$

[4]Алгоритм ТТ-округления контролирует относительную точность тензоров по норме Фробениуса, но для векторов норма Фробениуса совпадает с L_2-нормой.

Основной результат представлен в теореме 2, затем приведено следствие, которое легче интерпретировать.

Теорема 2. *Для любого набора факторов $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_m$ и любого значения параметра точности округления $\varepsilon \geq 0$ абсолютная ошибка оценки нормировочной константы, посчитанной алгоритмом 1, не превосходит:*

$$\begin{aligned}\left|Z - \widehat{Z}\right| \leq \|\boldsymbol{B}_1\|_2 \ldots \|\boldsymbol{B}_{n-2}\|_2 \cdot \|\boldsymbol{B}_{n-1}\boldsymbol{f}_n - \boldsymbol{f}_{n-1}\|_2 + \\ + \|\boldsymbol{B}_1\|_2 \ldots \|\boldsymbol{B}_{n-3}\|_2 \cdot \|\boldsymbol{B}_{n-2}\boldsymbol{f}_{n-1} - \boldsymbol{f}_{n-2}\|_2 + \ldots + \\ + \|\boldsymbol{B}_1\boldsymbol{f}_2 - \boldsymbol{f}_1\|_2\end{aligned} \tag{12}$$

Следствие 1. *Для любого набора факторов $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_m$ и любого значения параметра точности округления $\varepsilon \geq 0$ абсолютная ошибка оценки нормировочной константы, посчитанной алгоритмом 1, не превосходит:*

$$\left|Z - \widehat{Z}\right| \leq \|\boldsymbol{B}_1\|_2 \ldots \|\boldsymbol{B}_n\|_2 \left((1+\varepsilon)^{n-1} - 1\right) \tag{13}$$

Оценка из следствия менее точная, но позволяет по требуемой итоговой точности выбрать достаточный ε.

Чтобы пользоваться результатами теоремы 2, необходимо вычислять 2-норму векторов и матриц в TT-формате. 2-норма вектора совпадает с нормой Фробениуса соответствующего тензора, поэтому значения $\|\boldsymbol{B}_i\boldsymbol{f}_{i+1} - \boldsymbol{f}_i\|_2$ легко вычисляются. Хотя подсчет 2-нормы матриц в TT-формате является вычислительно сложной задачей, 2-норму можно оценить сверху с помощью нормы Фробениуса или с помощью эмпирически более точной оценки, использующей структуру матрицы $\boldsymbol{B}_i$:

$$\begin{aligned}\|\boldsymbol{B}_i\|_2 = \left\|\sum_{x_i} \boldsymbol{G}_i^1[x_i] \otimes \ldots \otimes \boldsymbol{G}_i^m[x_i]\right\|_2 \leq \sum_{x_i} \left\|\boldsymbol{G}_i^1[x_i] \otimes \ldots \otimes \boldsymbol{G}_i^m[x_i]\right\|_2 = \\ = \sum_{x_i} \left\|\boldsymbol{G}_i^1[x_i]\right\|_2 \ldots \|\boldsymbol{G}_i^m[x_i]\|_2 = U_i.\end{aligned} \tag{14}$$

Здесь используется равенство: $\|\boldsymbol{A} \otimes \boldsymbol{B}\|_2 = \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_2$.

На рис. 3 представлены результаты сравнения величины нормы Фробениуса $\|\boldsymbol{B}_i\|_F$, спектральной нормы $\|\boldsymbol{B}_i\|_2$ и её верхней оценки U_i. Значения разных норм указаны для всех индексов $i = 1, \ldots, n$ фиксированной модели Изинга.

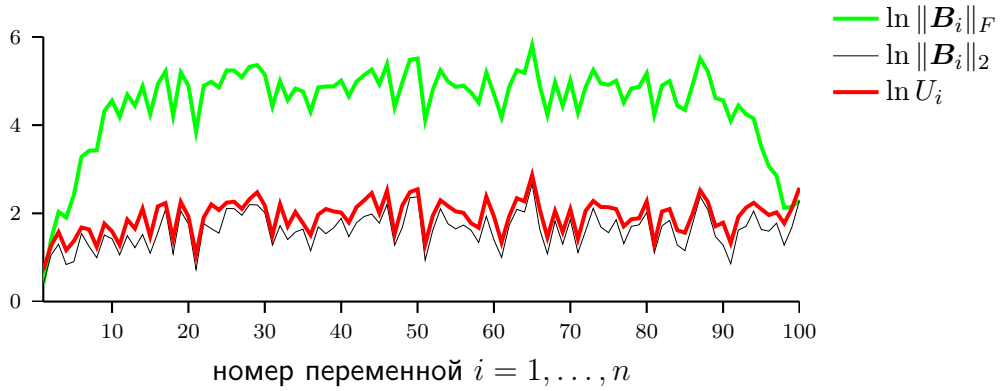

Рис. 3: Сравнение Фробениусовой и спектральной нормы матрицы $\boldsymbol{B}_i$ с верхней оценкой U_i. График построен для модели Изинга с решеткой размера 10×10, в которой коэффициенты унарных и парных потенциалов сгенерированы из равномерного распределения на $[-1, 1]$, а температура равна 1.

6.3 Доказательство теоремы 2

Лемма 1. *В условиях теоремы 2 для всех индексов $i = 1, \ldots, n-1$ верно следующее неравенство:*

$$\begin{aligned} &\|\boldsymbol{B}_i \ldots \boldsymbol{B}_n - \boldsymbol{f}_i\|_2 \leq \|\boldsymbol{B}_i\|_2 \ldots \|\boldsymbol{B}_{n-2}\|_2 \cdot \|\boldsymbol{B}_{n-1}\boldsymbol{f}_n - \boldsymbol{f}_{n-1}\|_2 + \\ &+ \|\boldsymbol{B}_i\|_2 \ldots \|\boldsymbol{B}_{n-3}\|_2 \cdot \|\boldsymbol{B}_{n-2}\boldsymbol{f}_{n-1} - \boldsymbol{f}_{n-2}\|_2 + \ldots + \|\boldsymbol{B}_i\boldsymbol{f}_{i+1} - \boldsymbol{f}_i\|_2 . \end{aligned} \tag{15}$$

Доказательство. Проведем доказательство по индукции.

В качестве базы индукции рассмотрим $i = n - 1$:

$$\|\boldsymbol{B}_{n-1}\boldsymbol{B}_n - \boldsymbol{f}_{n-1}\|_2 = \|\boldsymbol{B}_{n-1}\boldsymbol{f}_n - \boldsymbol{f}_{n-1}\|_2 .$$

(т. к. $\boldsymbol{f}_n = \boldsymbol{B}_n$ по построению.)

Предположим теперь, что (15) верно $\forall i = j+1, \ldots, n-1$. Тогда для $i = j$ получаем

$$\begin{aligned} &\|\boldsymbol{B}_j \ldots \boldsymbol{B}_n - \boldsymbol{f}_j\|_2 = \\ &= \|(\boldsymbol{B}_j \ldots \boldsymbol{B}_n - \boldsymbol{B}_j\boldsymbol{f}_{j+1}) + (\boldsymbol{B}_j\boldsymbol{f}_{j+1} - \boldsymbol{f}_j)\|_2 \leq \\ &\leq \|\boldsymbol{B}_j\|_2 \|\boldsymbol{B}_{j+1} \ldots \boldsymbol{B}_n - \boldsymbol{f}_{j+1}\|_2 + \|\boldsymbol{B}_j\boldsymbol{f}_{j+1} - \boldsymbol{f}_j\|_2 \leq \\ &\leq \|\boldsymbol{B}_j\|_2 (\|\boldsymbol{B}_{j+1}\|_2 \ldots \|\boldsymbol{B}_{n-2}\|_2 \cdot \|\boldsymbol{B}_{n-1}\boldsymbol{f}_n - \boldsymbol{f}_{n-1}\|_2 + \\ &+ \|\boldsymbol{B}_{j+1}\|_2 \ldots \|\boldsymbol{B}_{n-3}\|_2 \cdot \|\boldsymbol{B}_{n-2}\boldsymbol{f}_{n-1} - \boldsymbol{f}_{n-2}\|_2 + \\ &+ \ldots + \|\boldsymbol{B}_{j+1}\boldsymbol{f}_{j+2} - \boldsymbol{f}_{j+1}\|_2) + \|\boldsymbol{B}_j\boldsymbol{f}_{j+1} - \boldsymbol{f}_j\|_2 . \end{aligned}$$

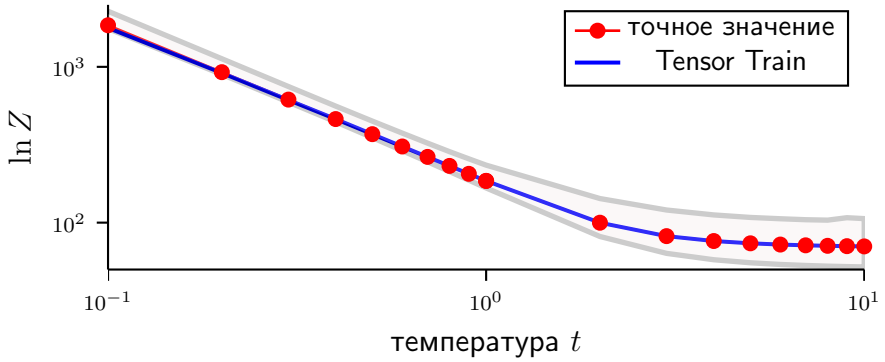

Рис. 4: Доверительный интервал на значение логарифма нормировочной константы $\ln Z$, полученный из теоремы 2 и неравенства (14). Детали см. в разделе 10.1.

Таким образом, (15) верно для $i = j$. □

Доказательство теоремы 2. По построению, $Z = \boldsymbol{B}_1 \dots \boldsymbol{B}_n$, а $\widehat{Z} = \boldsymbol{f}_1$. Таким образом, выполнены равенства

$$|Z - \widehat{Z}| = |\boldsymbol{B}_1 \dots \boldsymbol{B}_n - \boldsymbol{f}_1| = \|\boldsymbol{B}_1 \dots \boldsymbol{B}_n - \boldsymbol{f}_1\|_2 .$$

Здесь используется тот факт, что $\boldsymbol{B}_1 \dots \boldsymbol{B}_n$ и $\boldsymbol{f}_1$ — числа, а для чисел модуль совпадает с L_2-нормой одноэлементного вектора. Для завершения доказательства осталось применить лемму 1 к полученному выражению. □

6.4 Доказательство следствия 1

Лемма 2. *В условиях следствия 1 для всех индексов $i = 1, \dots, n$ верно следующее неравенство:*

$$\|\boldsymbol{f}_i\|_2 \le \|\boldsymbol{B}_i\|_2 \dots \|\boldsymbol{B}_n\|_2 \, (1+\varepsilon)^{n-i}. \tag{16}$$

Доказательство. Проведем доказательство по индукции.

Для $i = n$ утверждение следует из определения $\boldsymbol{f}_n$.

Пусть (16) верно $\forall j = i+1, \dots, n$. Тогда из (11) следует, что

$$\begin{aligned}\|\boldsymbol{f}_j\|_2 &= \|\boldsymbol{f}_j - \boldsymbol{B}_j \boldsymbol{f}_{j+1} + \boldsymbol{B}_j \boldsymbol{f}_{j+1}\|_2 \le \\ &\le \varepsilon \, \|\boldsymbol{B}_j\|_2 \, \|\boldsymbol{f}_{j+1}\|_2 + \|\boldsymbol{B}_j\|_2 \, \|\boldsymbol{f}_{j+1}\|_2 = \\ &= \|\boldsymbol{B}_j\|_2 \, \|\boldsymbol{f}_{j+1}\|_2 \, (1+\varepsilon).\end{aligned}$$

Чтобы завершить доказательство, остаётся воспользоваться предположением индукции для $\|\boldsymbol{f}_{j+1}\|_2$. □

Лемма 3. *В условиях следствия 1 для всех индексов $i = 1, \ldots, n-1$ верно следующее неравенство:*

$$\|\boldsymbol{B}_i \boldsymbol{f}_{i+1} - \boldsymbol{f}_i\|_2 \leq \|\boldsymbol{B}_i\|_2 \ldots \|\boldsymbol{B}_n\|_2 \, \varepsilon (1+\varepsilon)^{n-i-1}. \tag{17}$$

Доказательство. Утверждение следует из леммы 2 и неравенства

$$\|\boldsymbol{B}_i \boldsymbol{f}_{i+1} - \boldsymbol{f}_i\|_2 \leq \varepsilon \, \|\boldsymbol{B}_i \boldsymbol{f}_{i+1}\|_2 \leq \varepsilon \, \|\boldsymbol{B}_i\|_2 \, \|\boldsymbol{f}_{i+1}\|_2 \,,$$

которое в свою очередь следует из (11). □

Доказательство следствия 1. Применяя лемму 3 к неравенству (12), получаем

$$\begin{aligned} |Z - \widehat{Z}| \leq \|\boldsymbol{B}_1\|_2 \ldots \|\boldsymbol{B}_n\|_2 \, \varepsilon + \|\boldsymbol{B}_1\|_2 \ldots \|\boldsymbol{B}_n\|_2 \, \varepsilon (1+\varepsilon) + \ldots + \\ + \|\boldsymbol{B}_1\|_2 \ldots \|\boldsymbol{B}_n\|_2 \, \varepsilon (1+\varepsilon)^{n-2} = \|\boldsymbol{B}_1\|_2 \ldots \|\boldsymbol{B}_n\|_2 \, \varepsilon (1 + (1+\varepsilon) + \ldots + (1+\varepsilon)^{n-2}). \end{aligned}$$

Пользуясь формулой суммы геометрической прогрессии $1 + (1+\varepsilon) + \ldots + (1+\varepsilon)^{n-2} = ((1+\varepsilon)^{n-1} - 1)/\varepsilon$, получаем (13). □

7 Метод n-окрестностей

В этом разделе будет приведен обзор метода n-окрестностей [10]. Пусть задана модель Изинга, т. е. графическая модель, чья функция энергии следующим образом выражается через параметры $\boldsymbol{T} \in \mathbb{R}^{n \times n}$ и $\boldsymbol{h} \in \mathbb{R}^{n \times 1}$:

$$E(\boldsymbol{x}) = -\frac{1}{2} \sum_{i,j} T_{ij} x_i x_j - \sum_{i=1}^{n} h_i x_i, \quad x_i \in \{-1, 1\} \quad \forall i = 1 \ldots n, \tag{18}$$

где $\boldsymbol{x}$ – это вектор переменных модели, $\boldsymbol{T}$ – симметричная матрица параметров с нулями на диагонали и $\boldsymbol{h}$ – n-мерный вектор параметров. Положим $\boldsymbol{x}_0$ – конфигурацию минимума энергии, т. е. $\boldsymbol{x}_0 = \operatorname{argmin}. E(\boldsymbol{x})$. Запишем определение нормировочной константы:

$$Z = \sum_{\boldsymbol{x}} \exp(-E(\boldsymbol{x})) = \sum_{k=0}^{n} |\mathcal{X}^k_{\text{layer}}| \sum_{\boldsymbol{x} \in \mathcal{X}^k_{\text{layer}}} \frac{1}{|\mathcal{X}^k_{\text{layer}}|} \exp(-E(\boldsymbol{x})),$$

где $\mathcal{X}^k_{\text{layer}}$ – это множество, состоящее из всех наборов $\boldsymbol{x}$, которые отличаются от $\boldsymbol{x}_0$ ровно в k позициях. Множество $\mathcal{X}^k_{\text{layer}}$ называется *k-ым слоем*. Легко показать, что мощность k-ого слоя равна биномиальному коэффициенту: $|\mathcal{X}^k_{\text{layer}}| = \binom{n}{k}$.

Рассмотрим равномерное распределение вероятности на множестве $\mathcal{X}^k_{\text{layer}}$:

$$p_k(\boldsymbol{x}) = \begin{cases} \frac{1}{|\mathcal{X}^k_{\text{layer}}|}, & \text{если } \boldsymbol{x} \in \mathcal{X}^k_{\text{layer}} \\ 0, & \text{иначе.} \end{cases}$$

Тогда:

$$\mathbb{E}_{p_k(\boldsymbol{x})} \exp(-E(\boldsymbol{x})) = \sum_{\boldsymbol{x} \in \mathcal{X}^k_{\text{layer}}} \frac{1}{|\mathcal{X}^k_{\text{layer}}|} \exp(-E(\boldsymbol{x})) \tag{19}$$

Обращаем внимание, что компоненты случайной величины $\boldsymbol{x} \sim p_k$ не являются независимыми (например последняя компонента вектора $\boldsymbol{x} \in \mathcal{X}^k_{\text{layer}}$ однозначно определяется по первым $n-1$).

Рассмотрим теперь распределение на значения энергий: $q_k(e) = \frac{|\{\boldsymbol{x} \,|\, \boldsymbol{x} \in \mathcal{X}^k_{\text{layer}}, E(\boldsymbol{x}) = e\}|}{|\mathcal{X}^k_{\text{layer}}|}$. Тогда:

$$\mathbb{E}_{q_k(E)} \exp(-E) = \sum_{E \,\in\, \{E(\boldsymbol{x}) \,|\, \boldsymbol{x} \in \mathcal{X}^k_{\text{layer}}\}} q_k(E) \exp(-E) = \sum_{\boldsymbol{x} \in \mathcal{X}^k_{\text{layer}}} \frac{1}{|\mathcal{X}^k_{\text{layer}}|} \exp(-E(\boldsymbol{x}))$$

Случайная величина E – это сумма большого числа слагаемых вида $T_{ij} x_i x_j$, большинство из которых являются независимыми. Приблизим $q_k(E)$ нормальным распределением[5]. Тогда:

$$\sum_{\boldsymbol{x} \in \mathcal{X}^k_{\text{layer}}} \frac{1}{|\mathcal{X}^k_{\text{layer}}|} \exp(-E(\boldsymbol{x})) = \sum_{E \in \{E(\boldsymbol{x}) \,|\, \boldsymbol{x} \in \mathcal{X}^k_{\text{layer}}\}} q_k(E) \exp(-E) \approx \int_{E_{\min}}^{E_{\max}} f_k(E) \exp(-E) dE,$$

где $f_k(E)$ – плотность нормального распределения: $E \sim \mathcal{N}(E \mid \mu_k, \sigma_k^2)$, $E_{\min}$ и $E_{\max}$ – минимальное и максимальное значения функции энергии соответственно. Этот прием является ключевым для метода n-окрестностей.

Для поиска параметров нормального распределения $f_k(E)$, наилучшим образом приближающего распределение $q_k(E)$, приравняем моменты: $\mu_k = \mathbb{E}_{q_k} E$, $\sigma_k^2 = \mathbb{D}_{q_k} E$. Эти величины рассчитывается аналитически [10]. Пример выражения для среднего:

$$\mu_k = -\frac{\boldsymbol{x}_0^\intercal \boldsymbol{T} \boldsymbol{x}_0}{2} \frac{(n-2k)^2 - n}{n(n-1)} - (1 - \frac{2k}{n}) \boldsymbol{x}_0^\intercal \boldsymbol{h}$$

[5]Отметим, что $E \sim q_k$ – это дискретная случайная величина, и нормальное распределение является довольно грубым приближением для неё.

В ходе дальнейших преобразований[6] приходим к следующему приближенному выражению для нормировочной константы:

$$\begin{aligned} Z &= \sum_{k=0}^{n} \binom{n}{k} \sum_{\boldsymbol{x} \in \mathcal{X}^k_{\text{layer}}} \frac{1}{|\mathcal{X}^k_{\text{layer}}|} \exp(-E(\boldsymbol{x})) \approx \\ &\approx n \frac{1}{\sqrt{2\pi}} \int_0^1 \exp(nf(x)) \ \left(\Phi(B_x) - \Phi(A_x)\right) dx, \end{aligned} \tag{20}$$

где

$$\begin{aligned} f(x) &= -x \ln(x) - (1-x)\ln(1-x) - \frac{\mu_k}{n} + \frac{1}{2}\frac{\sigma_k^2}{n}, \\ A_x &= \sqrt{n}\left(\frac{E_{\min} - \mu}{\sigma} + \frac{\sigma}{n}\right), \\ B_x &= \sqrt{n}\left(\frac{E_{\max} - \mu}{\sigma} + \frac{\sigma}{n}\right). \end{aligned}$$

7.1 Ограничения подхода

Рассмотрим особенности метода, ограничивающие его применение в области графических моделей:

1. Ограниченность модели. В большинстве прикладных задач возникающие марковские случайные поля существенно выходят за рамки модели Изинга (многоклассовые модели, присутствие потенциалов высоких порядков, модели RBM, и т. п.).

2. Нарушение предположений метода. Т.к. энергия – это сумма *зависимых* слагаемых, то предположение о нормальности возникающего распределения нарушается на практике.

3. Скорость работы. Необходимость поиска минимума и максимума энергии, а так же численного взятия интеграла (20) ограничивает применимость метода к обучению MRF.

В разделе 8 предлагается изменение метода, направленное на устранение описанных недостатков.

[6] Основными преобразованиями являются: применение формулы Стирлинга, приближение внешней суммы интегралом и выделение полного квадрата под экспонентой.

8 Метод «Суммировать всё»

В этом разделе описывается предлагаемый метод, основанный на методе n-окрестностей и получивший название «Суммировать всё».

Рассмотрим произвольное марковское случайное поле. Предположим, что эмпирическая плотность распределения энергий E близка к нормальному распределению:

$$\forall y \quad P(E < y) = \frac{\sum_{\boldsymbol{x}}[E(\boldsymbol{x}) < y]}{|\mathcal{X}|} \approx \int_{-\infty}^{y} \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(t-\mu)^2}{2\sigma^2})}_{f(t \mid \mu,\sigma^2)} dt,$$

где $|\mathcal{X}|$ – это мощность совместной области определения всех переменных $x_1, \ldots, x_n$: $|\mathcal{X}| = \prod_{i=1}^{n} d_i$.

Заменим сумму в определении нормировочной константы на интеграл по все числовой оси:

$$\ln Z \approx \ln\left(|\mathcal{X}| \int_{-\infty}^{\infty} f(E \mid \mu, \sigma^2) \exp(-E) dE\right) = -\mu + \frac{\sigma^2}{2} + \ln(|\mathcal{X}|). \tag{21}$$

Найдём аналитические выражения для μ и σ^2, рассчитанные по методу максимального правдоподобия для множества всех энергий $\{E(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}\}$:

$$\begin{aligned}
\mu &= \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x}} E(\boldsymbol{x}) = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x}} \sum_{\ell} \Theta_\ell(\boldsymbol{x}^\ell) \\
&= \frac{1}{|\mathcal{X}|} \sum_{\ell} \sum_{\boldsymbol{x}} \Theta_\ell(\boldsymbol{x}^\ell) = \frac{1}{|\mathcal{X}|} \sum_{\ell} \frac{|\mathcal{X}|}{|\mathcal{X}^\ell|} \sum_{\boldsymbol{x}^\ell} \Theta_\ell(\boldsymbol{x}^\ell) \\
&= \sum_{\ell} \frac{1}{|\mathcal{X}^\ell|} \sum_{\boldsymbol{x}^\ell} \Theta_\ell(\boldsymbol{x}^\ell),
\end{aligned} \tag{22}$$

где $|\mathcal{X}^\ell|$ – это мощность совместной области определения тех переменных x_i, от которых существенно зависит ℓ-ый потенциал.

Т.е. μ – это по сути взвешенная сумма всех значений всех потенциалов. Асимптотическая сложность расчёта данной величины линейна по количества параметров марковского случайного поля.

Получим теперь формулу для параметра дисперсии σ^2:

$$\begin{aligned}
\sigma^2 &= \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x}} (E(\boldsymbol{x}))^2 - \mu^2 = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x}} \left(\sum_{\ell} \Theta_\ell(\boldsymbol{x}^\ell)\right)^2 - \mu^2 \\
&= \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x}} \sum_{\ell_1,\ell_2} \Theta_{\ell_1}(\boldsymbol{x}^{\ell_1}) \Theta_{\ell_2}(\boldsymbol{x}^{\ell_2}) - \mu^2 \\
&= \frac{1}{|\mathcal{X}|} \sum_{\ell_1,\ell_2} \sum_{\boldsymbol{x}} \Theta_{\ell_1}(\boldsymbol{x}^{\ell_1}) \Theta_{\ell_2}(\boldsymbol{x}^{\ell_2}) - \sum_{\ell_1,\ell_2} \left(\frac{1}{|\mathcal{X}^{\ell_1}|} \sum_{\boldsymbol{x}^{\ell_1}} \Theta_{\ell_1}(\boldsymbol{x}^{\ell_1})\right) \left(\frac{1}{|\mathcal{X}^{\ell_2}|} \sum_{\boldsymbol{x}^{\ell_2}} \Theta_{\ell_2}(\boldsymbol{x}^{\ell_2})\right)
\end{aligned}$$

Для любой пары потенциалов ℓ_1, ℓ_2 введем следующие обозначения:

$$\boldsymbol{x}_{1,2}^{\ell_1} = \boldsymbol{x}^{\ell_1} \setminus \boldsymbol{x}^{\ell_2},$$
$$\boldsymbol{x}_{1,2}^{\ell_2} = \boldsymbol{x}^{\ell_2} \setminus \boldsymbol{x}^{\ell_1},$$
$$\boldsymbol{x}_{1,2}^{\cup} = \boldsymbol{x}^{\ell_2} \cup \boldsymbol{x}^{\ell_1},$$
$$\boldsymbol{x}_{1,2}^{\cap} = \boldsymbol{x}^{\ell_2} \cap \boldsymbol{x}^{\ell_1}.$$

Так же обозначим совместную область определения переменных $\boldsymbol{x}_{1,2}^{\cup}$ за $\mathcal{X}_{1,2}^{\cup}$, а совместную область определения переменных $\boldsymbol{x}_{1,2}^{\cap}$ за $\mathcal{X}_{1,2}^{\cap}$. Введём дополнительное обозначение для суммы всех значений потенциала: $\tau^{\ell} = \sum_{\boldsymbol{x}^{\ell}} \Theta_{\ell}(\boldsymbol{x}^{\ell})$.

Тогда

$$\sigma^2 = \sum_{\ell_1,\ell_2} \left(\frac{1}{|\mathcal{X}_{1,2}^{\cup}|} \sum_{\boldsymbol{x}_{1,2}^{\ell_1},\boldsymbol{x}_{1,2}^{\ell_2},\boldsymbol{x}_{1,2}^{\cap}} \Theta_{\ell_1}(\boldsymbol{x}^{\ell_1})\Theta_{\ell_2}(\boldsymbol{x}^{\ell_2}) - \frac{\tau^{\ell_1}}{|\mathcal{X}^{\ell_1}|}\frac{\tau^{\ell_2}}{|\mathcal{X}^{\ell_2}|} \right). \tag{23}$$

Рассмотрим пару потенциалов ℓ_1, ℓ_2 таких, что они существенно зависят от не пересекающихся групп переменных: $\boldsymbol{x}_{1,2}^{\cap} = \emptyset$. Тогда $|\mathcal{X}_{1,2}^{\cup}| = |\mathcal{X}^{\ell_1}| \cdot |\mathcal{X}^{\ell_2}|$, и

$$\sum_{\boldsymbol{x}_{1,2}^{\ell_1},\boldsymbol{x}_{1,2}^{\ell_2},\boldsymbol{x}_{1,2}^{\cap}} \Theta_{\ell_1}(\boldsymbol{x}^{\ell_1})\Theta_{\ell_2}(\boldsymbol{x}^{\ell_2}) = \left(\sum_{\boldsymbol{x}^{\ell_1}} \Theta_{\ell_1}(\boldsymbol{x}^{\ell_1}) \right) \left(\sum_{\boldsymbol{x}^{\ell_2}} \Theta_{\ell_2}(\boldsymbol{x}^{\ell_2}) \right),$$

т. е. соответствующее слагаемое обнуляется. Таким образом, можно переписать выражение (23) добавив туда условие $\boldsymbol{x}_{1,2}^{\cap} \neq \emptyset$:

$$\begin{aligned} \sigma^2 &= \sum_{\substack{\ell_1,\ell_2:\\ \boldsymbol{x}_{1,2}^{\cap}\neq\emptyset}} \left(\frac{1}{|\mathcal{X}_{1,2}^{\cup}|} \sum_{\boldsymbol{x}_{1,2}^{\ell_1},\boldsymbol{x}_{1,2}^{\ell_2},\boldsymbol{x}_{1,2}^{\cap}} \Theta_{\ell_1}(\boldsymbol{x}^{\ell_1})\Theta_{\ell_2}(\boldsymbol{x}^{\ell_2}) - \frac{\tau^{\ell_1}}{|\mathcal{X}^{\ell_1}|}\frac{\tau^{\ell_2}}{|\mathcal{X}^{\ell_2}|} \right) \\ &= \sum_{\substack{\ell_1,\ell_2:\\ \boldsymbol{x}_{1,2}^{\cap}\neq\emptyset}} \frac{1}{|\mathcal{X}_{1,2}^{\cup}|} \left(\gamma^{\ell_1,\ell_2} - \frac{1}{|\mathcal{X}_{1,2}^{\cap}|}\tau^{\ell_1}\tau^{\ell_2} \right), \end{aligned} \tag{24}$$

где

$$\gamma^{\ell_1,\ell_2} = \sum_{\boldsymbol{x}_{1,2}^{\cap}} \left(\sum_{\boldsymbol{x}_{1,2}^{\ell_1}} \Theta_{\ell_1}(\boldsymbol{x}^{\ell_1}) \right) \left(\sum_{\boldsymbol{x}_{1,2}^{\ell_2}} \Theta_{\ell_2}(\boldsymbol{x}^{\ell_2}) \right).$$

Асимптотическая сложность вычисления σ^2 равна $O(\sum_{\ell=1}^{m} |\mathcal{X}^{\ell}|^2)$. В случае, когда все переменные принимают d значений ($d_i = d$, $i = 1, \ldots, n$), а все потенциалы существенно зависят от p переменных (в реальных задачах чаще всего $p \leq 3$), то оценка упрощается до $O(md^{2p})$.

Замечание: чтобы получить несмещенную выборочную оценку дисперсии, нужно поделить сумму квадратов отклонений на размер выборки минус один: $\overline{\sigma^2} = \frac{|\mathcal{X}|}{|\mathcal{X}|-1}\sigma^2$.

Отметим, что в отличие от большинства методов оценки нормировочной константы, точность данного метода возрастает при увеличении размера модели, т. к. увеличивается размер выборки $|\{E(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}\}|$, по которой вычисляются параметры нормального распределения.

8.1 Эффективные формулы для модели Изинга

Выведем формулы для μ и σ^2 для частного случая – когда рассматриваемое марковское случайное поле является моделью Изинга (см. раздел 7). В этом случае каждая переменная принимает два возможных значения $x_i \in \{-1, 1\}$, $i = 1, \ldots, n$, а MRF состоит из двух видов потенциалов: унарных $\Theta_i(x_i) = -h_i x_i$, $i = 1, \ldots, n$ и парных $\Theta_{i,j}(x_i, x_j) = -\frac{1}{2}T_{ij}x_i x_j$, $i, j = 1, \ldots, n$. Среднее значение всех потенциалов равно нулю:

$$\tau^i = \sum_{x_i} \Theta_i(x_i) = -h_i + h_i = 0,$$

$$\tau^{i,j} = \sum_{x_i, x_j} \Theta_{i,j}(x_i, x_j) = -2T_{ij} + 2T_{ij} = 0.$$

Подставляя средние значения потенциалов в формулу (22) получим, что $\mu = 0$.

Для подсчёта дисперсии воспользуемся формулой (24). Рассмотрим отдельно все пары потенциалов зависящие от пересекающихся подмножеств переменных.

Рассмотрим парный потенциал $\Theta_{i,j}(x_i, x_j)$ и унарный потенциал $\Theta_i(x_i)$. Для них множество общих переменных $\boldsymbol{x}^{\cap}_{1,2} = \{x_i\}$, множество $\boldsymbol{x}^{\ell_1}_{1,2} = \{x_j\}$, множество $\boldsymbol{x}^{\ell_2}_{1,2} = \varnothing$, мощность совместной области определения $|\mathcal{X}^{\cup}_{1,2}|$ равна 4.

$$\gamma^{(ij),i} = \sum_{x_i} \underbrace{\left(\sum_{x_j} \frac{1}{2} x_i x_j T_{ij} \right)}_{0} x_i h_i = 0.$$

Рассмотрим парные потенциалы $\Theta_{i,j}(x_i, x_j)$ и $\Theta_{i,k}(x_i, x_k)$, где $j \neq k$. Для них множество общих переменных $\boldsymbol{x}^{\cap}_{1,2} = \{x_i\}$, множество $\boldsymbol{x}^{\ell_1}_{1,2} = \{x_j\}$, множество $\boldsymbol{x}^{\ell_2}_{1,2} = \{x_k\}$.

$$\gamma^{(ij),(ik)} = \sum_{x_i} \underbrace{\left(\sum_{x_j} -\frac{1}{2} x_i x_j T_{ij} \right)}_{0} \underbrace{\left(\sum_{x_k} -\frac{1}{2} x_i x_k T_{ik} \right)}_{0} = 0.$$

Рассмотрим парные потенциалы $\Theta_{i,j}(x_i, x_j)$ и $\Theta_{i,j}(x_i, x_j)$. Для них множество общих переменных $\boldsymbol{x}_{1,2}^{\cap} = \{x_i, x_j\}$, множество $\boldsymbol{x}_{1,2}^{\ell_1} = \varnothing$, множество $\boldsymbol{x}_{1,2}^{\ell_2} = \varnothing$, мощность совместной области определения $|\mathcal{X}_{1,2}^{\cup}|$ равна 4.

$$\gamma^{(ij),(ij)} = \sum_{x_i, x_j} \frac{1}{2} x_i x_j T_{ij} \frac{1}{2} x_i x_j T_{ij} = T_{ij}^2.$$

Аналогично рассмотрим парные потенциалы $\Theta_{i,j}(x_i, x_j)$ и $\Theta_{j,i}(x_j, x_i)$: $\gamma^{(ij),(ji)} = T_{ij}^2$.

Рассмотрим унарные потенциалы $\Theta_i(x_i)$ и $\Theta_i(x_i)$. Для них множество общих переменных $\boldsymbol{x}_{1,2}^{\cap} = \{x_i\}$, множество $\boldsymbol{x}_{1,2}^{\ell_1} = \varnothing$, множество $\boldsymbol{x}_{1,2}^{\ell_2} = \varnothing$, мощность совместной области определения $|\mathcal{X}_{1,2}^{\cup}|$ равна 2.

$$\gamma^{i,i} = \sum_{x_i} x_i h_i x_i h_i = 2h_i^2.$$

Подставив выражения для γ^{ℓ_1,ℓ_2} в формулу (24) получим:

$$\sigma^2 = \frac{1}{2} \sum_{i,j} T_{ij}^2 + \sum_i h_i^2 = \frac{1}{2} \operatorname{Tr}(\boldsymbol{T}^\intercal \boldsymbol{T}) + \boldsymbol{h}^\intercal \boldsymbol{h}, \tag{25}$$

где под $\operatorname{Tr}(\boldsymbol{A})$ понимается след матрицы $\boldsymbol{A}$.

8.2 Эффективные формулы для модели RBM

Модель ограниченной машины Больцмана (Restricted Boltzmann Machines, RBM) [20] – это модель с бинарными переменными $x_i \in \{0, 1\}$, которые разбиваются на две группы: $\boldsymbol{x} = (v_1, \ldots, v_q, h_1, \ldots, h_f)$, где q – это число *наблюдаемых переменных*, а f – число *латентных переменных*, $q + f = n$. Функция энергии RBM задаётся следующим образом:

$$E(\boldsymbol{x}) = -\boldsymbol{b}^\intercal \boldsymbol{v} - \boldsymbol{c}^\intercal \boldsymbol{h} - \boldsymbol{h}^\intercal \boldsymbol{W} \boldsymbol{v},$$

где $\boldsymbol{b}$ и $\boldsymbol{c}$ – это q-мерный и f-мерный вектора параметров соотвественно, а $\boldsymbol{W}$ – это матрица параметров размера $f \times q$.

Применяя формулы (22) и (24) к потенциалам RBM получим:

$$\mu = -\frac{1}{4} \mathbf{1}^\intercal \boldsymbol{W} \mathbf{1} - \frac{1}{2} \boldsymbol{c}^\intercal \mathbf{1} - \frac{1}{2} \boldsymbol{b}^\intercal \mathbf{1},$$
$$\sigma^2 = \frac{1}{4}(\boldsymbol{c}^\intercal \boldsymbol{W} \mathbf{1} + \mathbf{1}^\intercal \boldsymbol{W} \boldsymbol{b}) + \frac{1}{16}(\mathbf{1}^\intercal \boldsymbol{W} \boldsymbol{W}^\intercal \mathbf{1} + \mathbf{1}^\intercal \boldsymbol{W}^\intercal \boldsymbol{W} \mathbf{1}) + \frac{1}{16} \operatorname{Tr}(\boldsymbol{W}^\intercal \boldsymbol{W}) + \frac{1}{4}(\boldsymbol{b}^\intercal \boldsymbol{b} + \boldsymbol{c}^\intercal \boldsymbol{c}).$$

9 Метод, основанный на формуле Тейлора

В этом разделе предлагается метод, не делающих предположения о нормальности распределения энергий.

Заменим в определении нормировочной константы (1) показательные функции на их разложение в ряд Тейлора до второго порядка вокруг точки E_0:

$$\begin{aligned} Z &= \sum_{\boldsymbol{x}} \exp(-E(\boldsymbol{x})) \approx \\ &\approx \sum_{\boldsymbol{x}} \left(\exp(-E_0) - \exp(-E_0)(E(\boldsymbol{x}) - E_0) + \exp(-E_0)\frac{1}{2}(E(\boldsymbol{x}) - E_0)^2 \right) \\ \ln Z &\approx \underbrace{-E_0 + \ln \left(\sum_{\boldsymbol{x}} 1 - (E(\boldsymbol{x}) - E_0) + \frac{1}{2}(E(\boldsymbol{x}) - E_0)^2 \right)}_{t(E_0)} \end{aligned}$$

Воспользовавшись определением μ (22) и σ^2 (24) можно упростить данное выражение:

$$\ln Z \approx t(E_0) = -E_0 + \ln \left(1 - \mu(1 + E_0) + \frac{1}{2}\mu^2 + \frac{1}{2}\sigma^2 + E_0 + \frac{1}{2}E_0^2 \right) + \ln(|\mathcal{X}|)$$

Отметим, что оценка логарифма нормировочной константы по метод «Суммировать всё» (21) является оценкой $t(\mu)$ через разложение в ряд Тейлора:

$$t(\mu) = -\mu + \ln \left(1 + \frac{1}{2}\sigma^2 \right) + \ln(|\mathcal{X}|) \approx \underbrace{-\mu + \frac{1}{2}\sigma^2 + \ln(|\mathcal{X}|)}_{\text{оценка по методу «Суммировать всё»}}$$

10 Вычислительные эксперименты

В этом разделе проводится экспериментальное сравнение метода n-окрестностей, предлагаемых методов «Суммировать всё», «Тейлор» и «Tensor Train», и альтернатив из библиотеки LibDAI [14]: Belief Propagation (BP) [11] и метод Mean Field (MF) [24]. Так же проводится сравнение с методом Annealed Importance Sampling (AIS) [15] — представителем методов MCMC.

Во всех экспериментах используется неоптимизированная реализация предложенных методов в среде Matlab[7]. Для операций связанных с TT-форматом используется TT-Toolbox[8], реализованный в среде Matlab.

[7] `https://github.com/bihaqo/TT-MRF`

[8] `http://spring.inm.ras.ru/osel/download/tt22.zip`

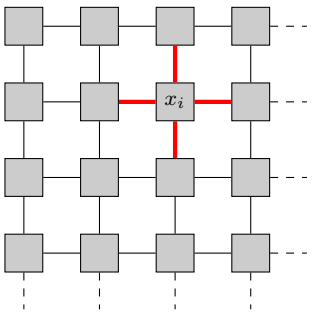

Рис. 5: 4-х связная решетка, задающая систему соседства: $(i, j) \in \mathcal{E}$ тогда и только тогда, когда переменные x_i и x_j соединены ребром.

10.1 Модель Изинга

Добавим параметр *температуры* $t \in \mathbb{R}_+$ в модель Изинга (см. раздел 7):

$$E(\boldsymbol{x}) = \frac{1}{t}\left(-\frac{1}{2}\sum_{i,j} T_{ij}x_i x_j - \sum_{i=1}^{N} h_i x_i\right).$$

Для каждого значения температуры t сгенерировано 50 моделей Изинга размера 10×10 со унарными весами сгенерированными из равномерного распределения $h_i \sim U[-1, 1]$, $i = 1, \ldots, n$ и парными весами равными 1 на 4-х связной решетке (рис. 5):

$$T_{ij} = \begin{cases} 1, & \text{если } (i, j) \in \mathcal{E} \\ 0, & \text{иначе.} \end{cases}$$

Размер 10×10 выбран достаточно большим, чтобы расчёт нормировочный константы по определению был невыполнимым (т. к. требует суммирования по $2^{10\times 10}$ слагаемым), но достаточно маленьким для того, чтобы можно было посчитать ошибку методов с помощью метода Junction Tree [24]. На рис. 6 указана медиана, а так же верхняя и нижняя квартили абсолютной ошибки оценки логарифма нормировочной константы. Медианное время работы методов указано в таблице 2.

На рис. 4 изображен доверительный интервал на значение логарифма нормировочной константы, полученный путем оценивания результата теоремы 2 неравенством (14).

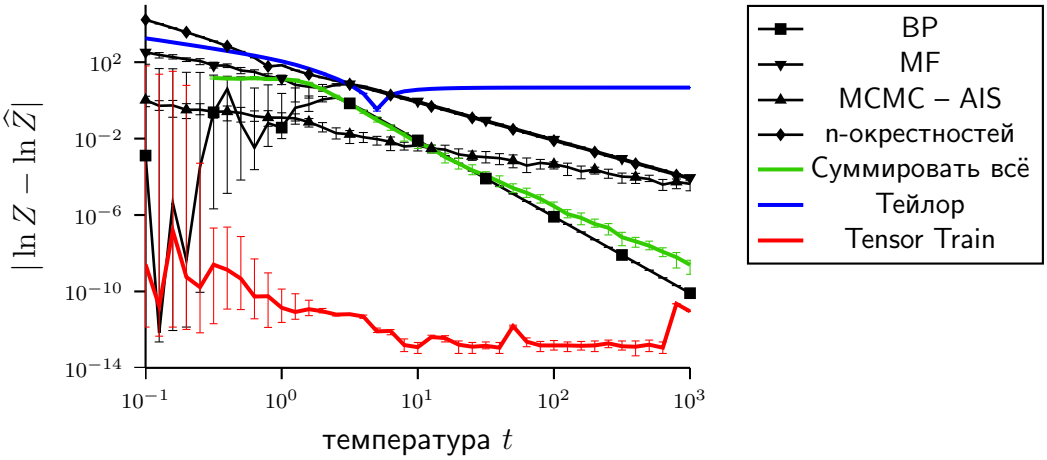

Рис. 6: Абсолютная ошибка подсчета логарифма нормировочной константы в зависимости от температуры (чем меньше, тем лучше).

10.2 TT-ранги энергии и вероятности

В данном эксперименте проверяется утверждение из разделов 5.1 и 5.2: при увеличении количества переменных TT-ранги тензора энергии растут медленно, а TT-ранги тензора ненормированной вероятности — экспоненциально быстро (рис. 2). Рассматривается последовательность моделей Изинга с 4-х связной решеткой возрастающих размеров: от 1×1 до 12×12. Унарные веса h_i сгенерированы из равномерного распределения $U[-1,1]$, все парные веса T_{ij} равны 1, температура t равна 10. Для моделей Изинга указан максимальный TT-ранг энергии и ненормированной вероятности при точном TT-представлении и после TT-округления с точностью 10^{-8}. Для точного TT-представления тензора вероятности модели Изинга с решеткой размера 11×11 не хватает 8GB оперативной памяти.

10.3 Обучение модели RBM

Исследуем точность рассматриваемых оценок логарифма нормировочной константы в процессе обучения модели RBM (см. раздел 8.2). В качестве обучающей выборки использовалась база данных MNIST [12] – черно-белые изображения рукописных цифр размера 28×28. Из 784 признаков (пикселей) были выбраны 250 с

Метод	Изинг, время (с)	RBM, время (с)
BP	0.005	509
MF	0.004	0.1147
MCMC – AIS	60	–
n-окрестностей	0.1505	–
«Суммировать всё»	0.00043	0.0041
«Тейлор»	0.0005	0.0047
«Tensor Train»	32	596

Таблица 2: Медианное время работы методов оценки нормировочной константы модели Изинга размера 10×10 и модели RBM с 250 наблюдаемыми и 12 латентными переменными.

наибольшей дисперсией [9]. Рассматривалась модель RBM с 12 латентными переменными (число 12 выбрано достаточно маленьким, чтобы был применим метод Junction Tree). Для обучения по методу максимального правдоподобия использовался метод наискорейшего подъёма:

$$\begin{aligned} \ln L(\boldsymbol{w}) &= \sum_{\ell} \Theta_\ell(\boldsymbol{T}_{\text{train}}; \boldsymbol{X}_{\text{train}}, \boldsymbol{w}) - \ln Z(\boldsymbol{X}_{\text{train}}, \boldsymbol{w}) \\ \alpha_e &= \arg\max_{\alpha} \ \ln L(\boldsymbol{w}^e + \alpha \nabla \ln L(\boldsymbol{w}^e)) \\ \boldsymbol{w}^{e+1} &= \boldsymbol{w}^e + \alpha_e \nabla \ln L(\boldsymbol{w}^e) \end{aligned}$$

Величина $\sum_\ell \Theta_\ell(\boldsymbol{T}_{\text{train}}; \boldsymbol{X}_{\text{train}}, \boldsymbol{w})$ и её градиент могут быть эффективно рассчитаны по определению, а для подсчёта логарифма нормировочной константы $\ln Z$ и её градиента использовался (точный) метод Junction Tree.

Для каждой из моделей RBM, заданных параметрами $\boldsymbol{w}^e$, $e = 1 \ldots Q$, вычислены оценки логарифма правдоподобия $\ln \widehat{L}(\boldsymbol{w}^e)$ и приведены на рис. 7 и 8. Медианное время работы методов приведено в таблице 2.

[9] Признаки, равные 0 (1) на большинстве объектов обучающей выборки, обладают априорной вероятностью близкой к 1 и вызывают численные неустойчивости.

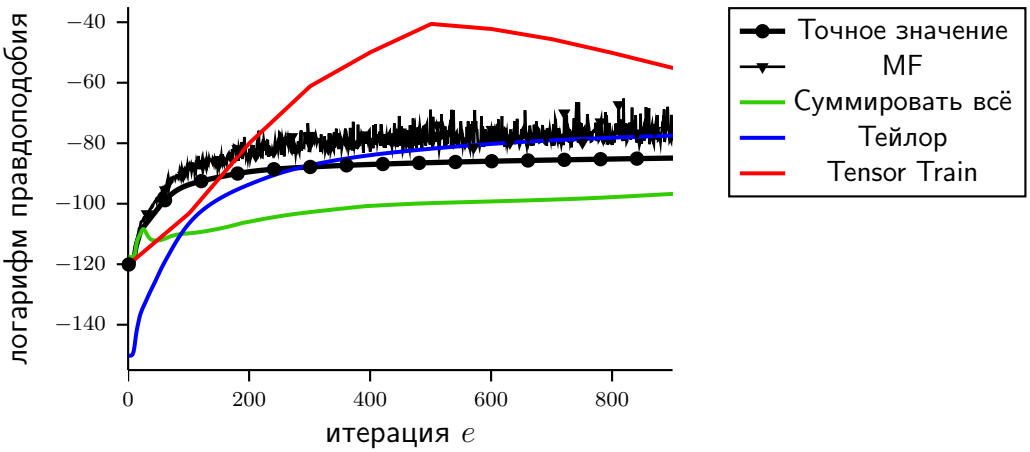

Рис. 7: Логарифм правдоподобия и его оценки на различных итерациях обучения модели RBM.

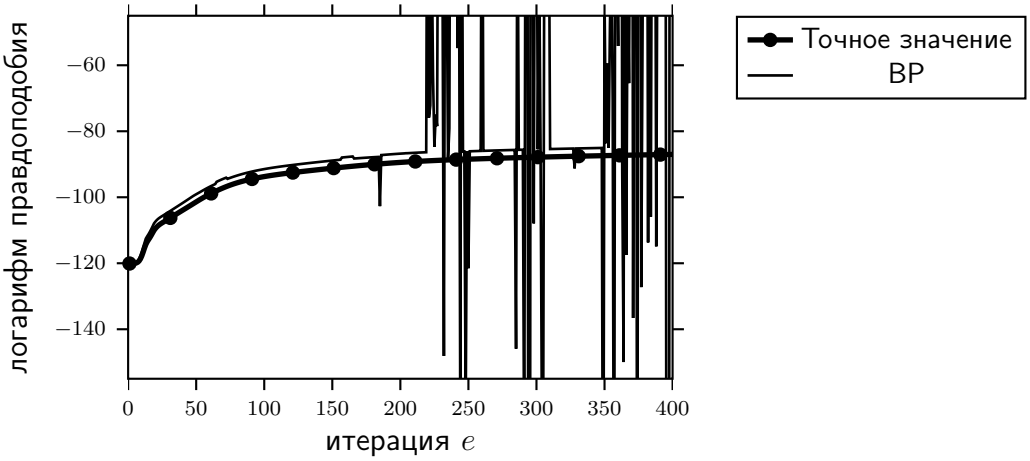

Рис. 8: Логарифм правдоподобия и его оценка методом Belied Propagation на различных итерациях обучения модели RBM.

10.4 Обсуждение и выводы

Предложенные методы «Суммировать всё» и метод, основанный на разложении в ряд Тейлора, на порядки опережают все аналоги по скорости работы (табл. 2). При этом метод «Суммировать всё» работает точнее метода n-окрестностей (рис. 6). В эксперименте по оценке логарифма нормировочной константы модели Изинга со случайными унарными потенциалами, метод, основанный на разложении Tensor Train, на порядки опережает все аналоги по точности работы (рис. 6). На первых итерациях обучения модели RBM (пока веса модели близки к случайной инициализации) метод «Tensor Train» так же показывает высокое качество работы, но по мере обу-

чения точность падает. В эксперименте по обучению RBM из всех рассмотренных методов только методы Mean Field и «Тейлор» показывают качественно верный результат (т. е. форма графика оценки повторяет форму графика точного логарифма правдоподобия), при этом метод Mean Field значительно менее устойчив по сравнению с предложенными методами (рис. 7). Метод Belief Propagation по мере обучения расходится (рис. 8).

11 Заключение

В настоящей дипломной работе получены следующие основные результаты:

- Предложен алгоритм оценки нормировочной константы через TT-разложение факторов графической модели, который не строит TT-представление тензора ненормированного совместного распределения;
- Приведены теоретические гарантии на точность оценки нормировочной константы;
- Найден практически важный класс задач, на котором тензорный метод опережает все доступные аналоги по точности работы;
- Предложено обобщение метода n-окрестностей, применимое к произвольным марковским случайным полям и обладающее более высокой скоростью работы;
- Предложен простой и быстрый метод оценки нормировочной константы через разложение в ряд Тейлора, который даёт качественно верную оценку на задаче обучения RBM.

По теме работы сделано 3 доклада на международных конференциях [17, 16, 19] и опубликована статья в журнале ВАК [25].

Список литературы

[1] J. D. Caroll and J. J. Chang. Analysis of individual differences in multidimensional scaling via n-way generalization of Eckart-Young decomposition. *Psychometrika*, 35:283–319, 1970.

[2] S. V. Dolgov and D. V. Savostyanov. Alternating minimal energy methods for linear systems in higher dimensions. Part II: Faster algorithm and application to nonsymmetric systems. arXiv preprint 1304.1222, 2013.

[3] L. Grasedyck. Hierarchical singular value decomposition of tensors. *SIAM J. Matrix Anal. Appl.*, 31:2029–2054, 2010.

[4] R. Grosse, C. Maddison, and R. Salakhutdinov. Annealing between distributions by averaging moments. In *Neural Information Processing Systems (NIPS)*, 2013.

[5] W. Hackbusch and S. Kühn. A new scheme for the tensor representation. *J. Fourier Anal. Appl.*, 15:706–722, 2009.

[6] T. Hazan, S. Maji, and T. Jaakkola. On sampling from the Gibbs distribution with random maximum a posteriori perturbations. In *Neural Information Processing Systems (NIPS)*, 2013.

[7] M. Ishteva, H. Park, , and L. Song. Unfolding latent tree structures using 4th order tensors. In *International Conference on Machine Learning (ICML)*, 2013.

[8] Y. Jernite, Y. Halpern, and D. Sontag. Discovering hidden variables in noisy-or networks using quartet tests. In *Neural Information Processing Systems (NIPS)*, 2013.

[9] B. N. Khoromskij and I. V. Oseledets. DMRG+QTT approach to computation of the ground state for the molecular Schrödinger operator. Preprint 69, MPI MIS, Leipzig, 2010.

[10] Boris Kryzhanovsky and Leonid Litinskii. Approximate method of free energy calculation for spin system with arbitrary connection matrix. *arXiv preprint arXiv:1410.6696*, 2014.

[11] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[12] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[13] T. Minka and Y. Qi. Tree-structured approximations by expectation propagation. In *Neural Information Processing Systems (NIPS)*, 2004.

[14] J. M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173, August 2010.

[15] R. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

[16] Alexander Novikov, Anton Rodomanov, Anton Osokin, and Dmitry Vetrov. Computationally efficient methods for map-inference and partition function estimation in mrf in tt format. In *SIAM Conference on Imaging Science*, 2014.

[17] Alexander Novikov, Anton Rodomanov, Anton Osokin, and Dmitry Vetrov. Putting mrfs on a tensor train. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 811–819, 2014.

[18] I. V. Oseledets. Tensor-Train decomposition. *SIAM J. Scientific Computing*, 33(5):2295–2317, 2011.

[19] Anton Rodomanov, Alexander Novikov, Anton Osokin, and Dmitry Vetrov. Low-rank approximation of energies in markov random fields and their representation in tt-format. In *SIAM Conference on Imaging Science*, 2014.

[20] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. 1986.

[21] L. Song, M. Ishteva, A. Parikh, E. Xing, and H. Park. Hierarchical tensor decomposition of latent tree graphical models. In *International Conference on Machine Learning (ICML)*, 2013.

[22] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[23] M. J. Wainwright, T. Jaakkola, and A. S. Willsky. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51(7):2313–2335, 2005.

[24] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[25] Александр Новиков, Антон Родоманов, Антон Осокин, and Дмитрий Ветров. Тензорный поезд в марковском случайном поле. *Журнал «Интеллектуальные системы. Терия и приложения»*, pages 293–318, 2014.