ФИЗИКА ЭЛЕМЕНТАРНЫХ ЧАСТИЦ И АТОМНОГО ЯДРА
2022. Т. 53, вып. 4. С. 904–930

ОТ СПИНОВЫХ СТЕКОЛ К ОБУЧЕНИЮ НЕЙРОННЫХ СЕТЕЙ

Е. Е. Перепёлкин [1,2,3], *Б. И. Садовников* [2],
Н. Г. Иноземцева [3,4], *Р. А. Рудаменко* [2],
А. А. Тарелкин [2], *П. Н. Сысоев* [2], *Р. В. Полякова* [1,*],
М. Б. Садовникова [2]

[1] Объединенный институт ядерных исследований, Дубна
[2] Московский государственный университет им. М. В. Ломоносова, Москва
[3] Московский технический университет связи и информатики, Москва
[4] Государственный университет «Дубна», Дубна, Россия

Дан обзор концептуальной основы теории спиновых стекол. Приведено описание математического аппарата, разработанного для спиновых стекол и модели ограниченной машины Больцмана (ОМБ). Рассмотрена оптимизация алгоритма обучения ОМБ неградиентными методами. Описан и использован способ извлечения гиперпараметра алгоритма обучения — температуры. Исследованы критические явления в ОМБ — энтропийный кризис, различие в температурах создания и обработки обучающей выборки.

This paper gives an overview of the conceptual basis of the theory of spin glasses. The mathematical approach developed for spin glasses and the model of the restricted Boltzmann machine (RBM) are described. The optimization of the RBM learning algorithm by non-gradient methods is considered. The method for extraction of the learning algorithm hyperparameter — temperature — is described and used. The critical phenomena in RBM — entropy crisis, difference in temperatures of training and testing sets — are explored.

PACS: 07.05.Mh; 42.79.Ta; 03.65.Yz; 05.30.–d; 67.10.Fj

ВВЕДЕНИЕ

Со времен появления вычислительной техники человек стремился создать машину, способную «мыслить» — решать задачи с неизвестным заранее набором условий.

В 1959 г. Артур Самуэль, исследователь искусственного интеллекта, ввел термин «машинное обучение». Он изобрел первую самообучающуюся компьютерную программу по игре в шашки. Самуэль определил машинное обучение как процесс, в результате которого компьютеры

* E-mail: polykovarv@mail.ru

способны показать такое поведение, которое в них не было запрограммировано изначально [1]. В основе его программы лежал рекурсивный алгоритм, часто используемый в принятии решений в теории игр. Этот алгоритм создает все пространство поиска игры для данной позиции, изображаемое в виде дерева возможных решений, и возвращает ход, связанный с наибольшим значением введенной оценочной величины (вознаграждения), независимо от хода соперника. Основным результатом его работы стали два метода, в которых использовалось обучение для создания одной из первых программ искусственного интеллекта: заучивание наизусть и обучение путем создания обобщений. Второй способ подразумевает работу со стохастическими алгоритмами, поэтому он стал основополагающим для концепции обучения машин.

Следующим этапом развития машинного обучения с ростом вычислительных возможностей [2, 3] стал переход к моделям с бóльшим количеством промежуточных вычислений — появление нейронных сетей и глубокого обучения. Существует концептуальная аналогия между нейронными сетями и хаотическими физическими системами. Отношение между вероятностью и гармоничностью системы в теории информации аналогично соотношению вероятности и энергии в статистической физике. Указанная аналогия позволяет использовать математические методы, разработанные для описания таких систем, для количественного анализа некоторых режимов работы нейронных сетей [5].

Исследования хаотичных физических систем и флуктуаций могут иметь широкий спектр применения [4]. Популярная физическая модель, используемая для работы с такими системами, — спиновые стекла. Спиновое стекло — это стохастично расположенные в пространстве магнитные моменты с различными видами взаимодействия. На микроскопическом уровне спиновые стекла состоят из множества элементарных спинов, беспорядочно взаимодействующих друг с другом посредством парных сил. По отдельности эти силы пытаются ориентировать пары спинов параллельно или антипараллельно, но в совокупности они приводят к фрустрации относительно глобальных ориентаций. Под фрустрацией (frustration) понимается ситуация неопределенности спина для группы из трех спинов при парном взаимодействии [4]. Следствием фрустрации является система со множеством неэквивалентных метастабильных глобальных состояний, а значит, множеством интересных физических свойств.

В настоящее время есть два аспекта, в которых анализ спинового стекла переходит в анализ нейронных сетей. Первый — исследование макроскопического асимптотического поведения нейронной сети с заданной архитектурой и весами связей. Второй касается выбора одних глобальных параметров для оптимизации других.

Задача выбора гиперпараметров играет существенную роль в машинном и глубоком обучении. Так, известные классические алгоритмы машинного обучения требуют для точной работы оптимального выбора

гиперпараметров и большого количества обучающих данных, в то время как человек может понять концепцию из нескольких событий [6]. Таким образом, предварительная подготовка данных для последующей тренировки на них глубоких нейронных сетей [8] увеличивает конечную точность алгоритма, выделяя начальную область в пространстве параметров, с которой начинается тонкая настройка [9]. Извлечение не заданных явно признаков из неразмеченных выборок называется обучением без учителя [7]. Несмотря на высокую значимость этой процедуры, существует мало теоретических работ, посвященных тому, как происходит процесс неконтролируемого (без учителя) обучения. Одна из причин заключается в том, что процесс обучения без учителя в глубокой нейронной сети, как правило, очень сложен. Понимание и описание механизма неконтролируемого обучения в элементарных моделях играет важную роль, так как при работе с глубокими сетями часто возникает серьезная проблема — потеря интерпретации созданных признаков, т. е. утрата физического смысла.

Машина Больцмана (МБ) — одна из базовых моделей нейронных сетей [10]. Благодаря способности выявлять скрытые внутренние представления и решать сложные задачи комбинаторики МБ применяется в машинном обучении и построении статистических закономерностей, в том числе для предварительной обработки и обобщения данных [11]. Машины Больцмана представляют собой нейронные сети с симметрично соединенными слоями, разделенными на две категории — «видимые» и «скрытые». Несмотря на успех в практических применениях, строгое математическое описание машин Больцмана остается сложной задачей. В исследованиях МБ коэффициенты связей на ребрах считаются закрепленными, а их распределение извлекается при обучении [13]. Рассмотрение МБ может быть проведено в рамках статистической механики [14]. Свойство симметричности матрицы связей определяет тесную связь МБ с физической моделью спиновых стекол.

Целью данной работы является физическое описание ограниченной машины Больцмана (ОМБ) [12], в которой существуют связи между нейронами различных слоев, но нет внутренних, и исследование режимов ее работы аналитическими и численными методами.

Работа имеет следующую структуру. В п. 1.1 рассматривается известная модель спиновых стекол, для которой вводятся понятие намагниченности и ее обобщения — перекрытия, объясняется использование трюка реплик как метода приближенного вычисления статистической суммы системы из большого числа подсистем, обосновывается необходимость введения функции свободной энергии как функции намагниченности. В п. 1.2 рассматриваются известные аналогии между классической термодинамикой и теорией информации по введению описательных функций систем, в частности энтропии и температуры.

Разд. 2 посвящен математическому описанию простейшей ограниченной машины Больцмана с единственным скрытым нейроном и процессов

ее обучения — алгоритму передачи сообщений и максимизации правдоподобия. В этом же разделе рассмотрена ОМБ с бинарными связями, эквивалентная двухстороннему спиновому стеклу, в котором две стороны содержат переменные различной природы: видимый слой состоит из бинарных изинговых спинов, скрытый слой — из реальных гауссовых. Для решения вычислительных задач машина проходит обучение, где ее параметры — пороговые значения активации нейрона θ и коэффициенты связей на ребрах ξ — стохастически изменяются по выбранным алгоритмам. После этого видимый слой инициализируется заданным состоянием и происходит эволюция системы к стационарному состоянию. В результате выходной слой представляет решение задачи.

В п. 3.1 приводится описание работы ОМБ и вычисление необходимой статистической суммы для входящего набора данных с точки зрения методов Монте-Карло для марковских процессов с помощью оператора переходов $T(x'|x)$ из состояния x в состояние x', распределения вероятности состояний марковских цепей $g^{(t)}(x)$. В п. 3.2, 3.3 вводится температура обучающей выборки, которая используется в стратегии темперирования для поиска целевого распределения данных.

В разд. 4 приведены исследования физических характеристик ограниченной машины Больцмана с помощью описанных ранее методов, в частности: зависимость энтропии, приходящейся на нейрон, от плотности данных, оценка точности аппроксимированного алгоритма передачи сообщений в сравнении с классическим аналогом, изменение скорости расчета перекрытия при изменении температуры обработки данных относительно температуры создания данных в зависимости от плотности данных, влияние размера сети на скорость работы алгоритма расчета перекрытия, измерение температуры обучающей выборки для ее различных размеров, нахождение оптимального размера обучающей выборки.

В заключении приведены полученные результаты моделирования и их интерпретация.

1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ

1.1. Из статистической механики. Основной целью статистической механики является описание макроскопических характеристик систем многих тел с помощью взаимодействий между элементами на микроскопическом уровне. Задача состоит в том, чтобы вычислить термодинамическое среднее (среднее по ансамблю) значение физической величины, используя распределение Гиббса–Больцмана:

$$P_{\mathrm{GB}}(\sigma) = \frac{\mathrm{e}^{-\beta H}}{Z}, \tag{1.1}$$

где $\beta = T^{-1}$, T — температура системы; Z — статистическая сумма системы.

Для разных агрегатных состояний (фаз) вещества с одной и той же химической формулой макроскопические свойства могут сильно отли-

чаться друг от друга, поскольку межмолекулярные взаимодействия существенно изменяют макроскопическое поведение в зависимости от температуры, давления и других внешних условий. Чтобы исследовать общий механизм таких резких изменений макроскопических состояний материалов, вводится модель Изинга — одна из простейших моделей взаимодействующих систем многих тел. Стандартная теория для описания общих черт фазовых переходов строится следующим образом: назовем множество V целых чисел от 1 до N, $V = \{1, \ldots, N\}$, $\{i\} : i = 1, \ldots, N$, решеткой, а ее элемент i — сайтом. Под сайтом будем понимать некоторую абстракцию, например реальный узел кристаллической решетки, пиксель цифрового изображения или, возможно, нейрон в нейронной сети. Каждому сайту ставится в соответствие переменная типа изингового спина σ_i. Спин Изинга характеризуется бинарным значением $\sigma_i = \{\pm 1\}$, и этот случай будет рассматриваться далее. На рис. 1 показана двумерная модель спинового стекла. В задаче магнетизма спин Изинга σ_i показывает, направлен ли микроскопический магнитный момент вверх или вниз.

Атомы в кристалле расположены в узлах решетки через равные промежутки. В спиновых стеклах положение атомов в пространстве случайно, поэтому в теоретических расчетах необходимо ввести распределение вероятностей для распределения атомов, а значит, спинов и, следовательно, констант парных взаимодействий между спинами. Важным моментом является то, что в стеклах кажущееся случайным расположение атомов не меняется с течением времени на другой набор случайных расположений — существует закаленный беспорядок. Случайность в положении сайтов атомов считается менее значимой для макроскопических свойств спиновых стекол по сравнению со случайностью во взаимодействиях (случайностью связей). Таким образом, предполагается, что переменные парных взаимодействий распределены случайно и независимо на каждой связи J.

Термин спиновое стекло подразумевает, что ориентация спинов имеет сходство с таким расположением атомов в стеклах: спины случайным образом заморожены в спиновых стеклах. Теория спиновых стекол ставит

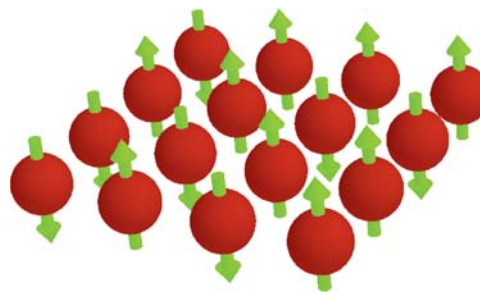

Рис. 1. Графическое представление решетки двумерного спинового стекла с изинговыми спинами

своей целью выяснение условий существования состояний спинового стекла. В рамках теории среднего поля установлено, что фаза спинового стекла существует при низких температурах, когда между спинами происходят случайные взаимодействия определенных типов. В спиновых стеклах беспорядок присутствует в гамильтониане в явном виде в форме сочетаний пар J по степеням свободы σ:

$$H = H(\sigma; J). \tag{1.2}$$

Беспорядок J полностью определяется распределением вероятностей $p(J)\,dJ$, одинаковым для каждой константы связи в системе. Известный пример — модель Эдвардса–Андерсона [15]

$$H = -\sum_{\langle ik \rangle} J_{ik}\sigma_i\sigma_k, \tag{1.3}$$

где спины σ_i — степени свободы, а связи J_{ik} — гауссовы случайные величины с явным видом распределения:

$$p(J_{ik}) = \frac{1}{\sqrt{2\pi J^2}} \exp\left\{-\frac{(J_{ik} - J_0)^2}{2J^2}\right\}, \tag{1.4}$$

со средней константой взаимодействия J_0 и дисперсией J^2.

Суммирование проводится по спинам ближайших соседей, т. е. модель конечномерна. «Закаленность» беспорядка означает, что J постоянны на всем временном интервале, на котором флуктуирует σ. Это будет важно, когда придется выполнять усреднение σ по J.

Даже если в гамильтониане нет закаленного беспорядка, при низкой температуре в застывшей конфигурации системы каждая частица испытывает действие неупорядоченной среды на себя. В этом смысле беспорядок самогенерируется. Причиной этого служит большое количество некристаллических локальных минимумов гамильтониана. Для работы с гамильтонианами типа (1.3) требуется избавиться от J, так как если бы от него зависела каждая пара спинов, было бы верно, что любое единичное изменение конфигурации изменяет физические свойства спинового стекла.

В соответствии с общей теорией статистической механики свободная энергия записывается в виде $F = E - TS$, где внутренняя энергия E — значение гамильтониана, а S — энтропия. Откуда следует, что свободная энергия пропорциональна логарифму статистической суммы Z:

$$F = -T\log Z = -T\log\,\mathrm{Tr}\,\mathrm{e}^{-H/T}. \tag{1.5}$$

Опишем систему в терминах свободной энергии, произведя функциональное интегрирование по ближайшим соседям (здесь $D\sigma = D[\sigma(i)]$, i — ближайший соседний спин):

$$F_N(J) = -\frac{1}{\beta N}\log\int \mathrm{e}^{-\beta H(\sigma;J)} D\sigma. \tag{1.6}$$

При устремлении числа спинов N в системе к бесконечности получаем

$$\lim_{N\to\infty} F_N(\beta, J) = F_\infty(\beta). \tag{1.7}$$

Свободная энергия больше не зависит от J и является функцией β. В таком случае усреднение по всем сочетаниям дает независимую от беспорядка величину

$$F = -\lim_{N\to\infty} \frac{1}{\beta N} \overline{\log Z(J)} = F_\infty(\beta), \tag{1.8}$$

где

$$\overline{\log Z(J)} = \int p(J) \log Z(J)\, dJ. \tag{1.9}$$

Это свойство называется самоусреднением. Суть самоусреднения в том, что распределение физических величин для больших N резко достигает пика вокруг их среднего значения, т. е. дисперсия их распределения стремится к нулю. Для вычисления свободной энергии необходим «трюк реплик»:

$$\overline{\log Z} = \lim_{n\to 0} \frac{1}{n} \log \overline{Z^n}. \tag{1.10}$$

В случае целого n справедливо выражение

$$\overline{Z^n} = \int \overline{\exp\left[-\beta H(\sigma_1; J) - \ldots - \beta H(\sigma_n; J)\right]} D\sigma_1 \cdots D\sigma_n. \tag{1.11}$$

Таким образом производим копирование системы («реплику») n раз, считаем все функцией n и устремляем n к нулю.

Одной из наиболее важных величин, используемых для характеристики макроскопических свойств модели, является намагниченность. Намагниченность с учетом усреднения (1.6) определяется как

$$m = \frac{1}{N} \sum_{i=1}^{N} \overline{\langle \sigma_i \rangle} \tag{1.12}$$

и измеряет общее упорядочение в макроскопической системе (т. е. в термодинамическом пределе $N \to \infty$). Намагниченность является мерой нахождения макроскопической системы в упорядоченном состоянии. Намагниченность исчезает, если существует равное количество направленных вверх спинов ($\sigma_i = 1$) и направленных вниз спинов ($\sigma_i = -1$), что указывает на отсутствие равномерно упорядоченного состояния. При низких температурах $\beta \gg 1$ распределение Гиббса–Больцмана (1.1) предполагает, что низкоэнергетические состояния реализуются с гораздо большей вероятностью, чем высокоэнергетические.

Чтобы описать схожесть двух конфигураций, используем понятие перекрытия, являющееся обобщением намагниченности:

$$q_{\sigma\tau} = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \tau_i, \tag{1.13}$$

где σ и τ — две различные конфигурации. Для изинговых спинов получаем

$$q_{\sigma\tau} = \begin{cases} 1, & \text{если } \sigma \text{ и } \tau \text{ почти совпадают,} \\ -1, & \text{если } \sigma \text{ и } \tau \text{ антикоррелированы,} \\ 0, & \text{если } \sigma \text{ и } \tau \text{ декоррелированы.} \end{cases} \tag{1.14}$$

В случае нарушения гипотезы эргодичности при низких температурах и больших N мера Гиббса распадается на составляющие, называемые чистыми состояниями

$$\langle * \rangle = \sum_{\alpha} \omega_\alpha \langle * \rangle_\alpha, \tag{1.15}$$

для которых перекрытие вводится как

$$q_{\alpha\beta} = \frac{1}{N} \sum_{i=1}^{N} \langle \sigma_i \rangle_\alpha \langle \sigma_i \rangle_\beta, \tag{1.16}$$

что может быть записано в виде

$$q_{\alpha\beta} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{Z_\alpha} \int\limits_{\sigma\in\alpha} \sigma_i \, \mathrm{e}^{-\beta H(\sigma)} D\sigma \frac{1}{Z_\beta} \int\limits_{\tau\in\beta} \tau_i \, \mathrm{e}^{-\beta H(\tau)} D\tau =$$

$$= \frac{1}{Z_\alpha Z_\beta} \int\limits_{\tau\in\beta} \int\limits_{\sigma\in\alpha} \mathrm{e}^{-\beta H(\sigma)} \, \mathrm{e}^{-\beta H(\tau)} q_{\sigma\tau} D\sigma \, D\tau. \tag{1.17}$$

Важной особенностью чистых состояний является свойство кластеризации. Это свойство показывает, что статистическая корреляция между двумя различными точками стремится к нулю, когда расстояние между ними стремится к бесконечности:

$$\langle \sigma_i \sigma_k \rangle \to \langle \sigma_i \rangle \langle \sigma_k \rangle, \quad |i-k| \to \infty. \tag{1.18}$$

В статическом рассмотрении спиновых стекол нарушение симметрии реплик происходит из-за нарушения эргодичности при критически низких температурах T_s. При динамическом подходе эргодичность нарушается при температуре $T_d > T_s$ из-за наличия метастабильных состояний, где T_d и T_s — температуры критических переходов для динамической и статической системы соответственно [16]. В частности, это приводит к получению отрицательной энтропии в процессе моделирования (см. разд. 4). Решение проблемы асимметричности реплик было предложено Дж. Паризи в виде итеративного разбиения диагональных элементов матрицы перекрытий и учета энергий отдельных блоков разбиений [17, 18].

Иную точку зрения на решение данной проблемы дает уравнение состояния, обусловленное локальной намагниченностью в спиновых стеклах. Для того чтобы указать отдельные состояния, удерживающие динамику, нужно ввести свободную энергию Таулесса–Андерсона–Палмера (ТАП). Чистые состояния — объекты, существующие в N-мерном фазо-

вом пространстве: в каждом состоянии α локальная намагниченность m имеет определенную величину, зависящую от узла $i : m_i^\alpha = \langle \sigma_i \rangle$, а состояние определяется вектором его намагниченностей. Поэтому необходима функция, заданная на этом пространстве, т. е. функция локальных намагниченностей m_i, локальные минимумы которой совпадают с чистыми состояниями системы. Минимизация такой функции должна обеспечить набор уравнений для вектора m_i, эквивалентный уравнению для m в бесконечномерной модели Изинга $m = \operatorname{th}(\beta m)$. Эта функция представляет собой среднюю свободную энергию, которая известна в теории спиновых стекол как свободная энергия ТАП. Важно подчеркнуть, что свободная энергия ТАП является функцией намагниченности m_i, а не микроскопических степеней свободы σ_i. В частности, ее минимумы не обязательно совпадают с минимумами энергии — минимумами гамильтониана $H(\sigma_i)$.

Чистые состояния нельзя просто отождествить с минимумами энергии из-за того, что различные минимумы энергии могут быть разделены энергетическими барьерами, которые при высокой температуре малы по сравнению с $k_B T$ и относятся к одному и тому же чистому состоянию. Даже при $T \to 0$, когда состояние системы эволюционирует до своей низшей энергетической конфигурации, важно различать эти понятия. Чистое состояние α, определяемое вектором $m_1^\alpha \cdots m_N^\alpha$, является подкомпонентом меры Гиббса.

Как было написано выше, чистое состояние обладает ключевым свойством кластеризации, которое не имеет смысла, когда речь идет о простой конфигурации $\sigma_1 \cdots \sigma_N$.

1.2. К топологической идентичности. Почему соотношение между вероятностью и гармоничностью системы в теории информации такое же, как и соотношение между вероятностью и энергией в статистической физике?

Из второго начала термодинамики для неравновесных процессов следует, что состоянием равновесия системы будет состояние с наибольшей энтропией при фиксированной средней энергии. Частицы будут занимать различные состояния, и макроскопические свойства системы будут зависеть от вероятности, с которой эти состояния будут заполнены. Энтропия такой системы является однозначной функцией плотности вероятностей и выражается формулой

$$S = -\sum_{i=1}^{n} P_i \ln P_i, \tag{1.19}$$

где P_i — вероятность реализации состояния i; n — количество возможных состояний [16]. Американский математик Клод Шеннон — отец-основатель теории информации — увидел, что однородность распределения, измеримая формулой для энтропии, является мерой потерь информации в распределении. Он показал, что экспоненциальное соотношение между гармоничностью системы и вероятностью следует из

максимизации потерь информации, точно так же, как в статистической механике оно выполняется для энергии и вероятности реализации данного состояния при максимизации энтропии. Таким образом, Шеннон ввел аналогию между информационной энтропией и физической. В дальнейшем аналогия была распространена на температуру.

В модели рассмотренных спиновых стекол каждый спин связан с остальными через взаимодействие. Такие связи могут быть представлены в виде полного графа. Рассмотрим расширение указанной модели до полных двудольных графов, которое назовем моделью двудольных спиновых стекол (ДСС) [17].

Основной мотивацией для изучения модели двудольных спиновых стекол является ее сходство с семейством нейронных сетей, известных как ограниченные машины Больцмана. ОМБ являются генеративными моделями, используемыми в обучении без учителя. Они привлекательны для изучения с точки зрения физики, поскольку представляют собой энергетические модели с гамильтонианом той же параметрической формы, что и модель ДСС. Фактически единственное реальное различие между этими моделями заключается в том, как выбираются параметры. В модели ДСС они распределяются нормально, тогда как в ОМБ они определяются алгоритмом обучения. Проводя анализ модели ДСС, можно получить новое понимание принципов работы ОМБ.

Модель двудольного спинового стекла описывается гамильтонианом вида

$$H = -\sum_{i=1}^{N}\sum_{k=1}^{P}\xi_{ik}\sigma_i h_k - \sum_{i=1}^{N} b_i^{(v)}\sigma_i - \sum_{k=1}^{P} b_k^{(h)} h_k, \tag{1.20}$$

где b_i — внешние поля; ξ — матрица взаимодействий; σ, h — состояния «видимого» и «скрытого» спинов; N — количество «видимых» спинов; P — количество «скрытых» спинов, введенных аналогично состояниям скрытой марковской цепи [8].

В данном случае считается, что явное взаимодействие происходит только между спинами «разных типов» — видимым и скрытым. Взаимодействия «видимый–видимый» и «скрытый–скрытый» описываются действием внешнего поля на каждый отдельно взятый спин. Именно этим обосновывается переход от полного графа к полному двудольному [17].

2. ПРИНЦИПЫ РАБОТЫ ОГРАНИЧЕННЫХ МАШИН БОЛЬЦМАНА

2.1. Определения и свойства. Желаемое поведение сети, определяемое данным числом входных («видимых») нейронов, достаточно часто может быть достигнуто простым подбором весов, соединяющих эти слои. Такое поведение демонстрирует сеть Хопфилда.

В большинстве случаев необходимо расширить архитектуру сети с помощью введения так называемых *скрытых* нейронов. Получается, что

скрытые нейроны не испытывают прямого влияния поступающей на вход информации. Таким образом, они позволяют уловить закономерности более высоких порядков, строя внутренние представления.

Машина Больцмана может рассматриваться как расширение сети Хопфилда — связи между нейронами в ней также двунаправленные, состояния представляют бинарное множество. В отличие от сети Хопфилда в машине Больцмана представлены скрытые нейроны, а переходы между состояниями задаются вероятностным распределением.

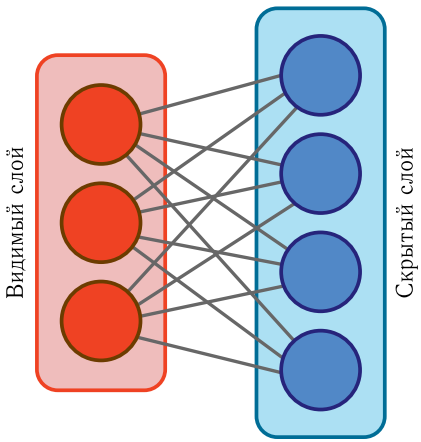

Рис. 2. Графическое представление ограниченной машины Больцмана

Ограниченная машина Больцмана состоит из двух слоев по аналогии с моделью двудольных спиновых стекол. Графическое представление приведено на рис. 2.

Работа ОМБ состоит из двух этапов:

1) обучение — процесс извлечения закономерностей из входных данных;

2) генерация — процесс создания (дополнения) данных, сгенерированных тем же распределением, что и входные данные, но отсутствующих в обучающей выборке.

Стандартным подходом к процессу обучения является метод сэмплирования (создания выборок) с применением градиентного спуска. Несмотря на успех в практических применениях, градиентный спуск не поддается аналитическим исследованиям [18]. Для возможности аналитического описания будем использовать байесовский вывод, учитывающий шум в данных и имеющий аналитическое обоснование на графических вероятностных моделях [19].

Рассмотрим обучение на неразмеченных данных (обучение без учителя) на конечномерных выборках для ОМБ с одним скрытым нейроном. Задачей обучения является выявление скрытого распределения данных ξ_i, поступающих на вход видимого слоя с N нейронами. Предположим, что компоненты вектора ξ_i равновероятно выбираются из множества $\{\pm 1\}$. Для упрощения рассмотрения исключим внешние поля, действующие на нейроны (смещения на языке машинного обучения). Энергия такой конфигурации определяется выражением

$$E = -\sum_{i=1}^{N} \xi_i \sigma_i h, \tag{2.1}$$

где σ — конфигурация видимого слоя, порожденная множеством $\{\pm 1\}$; h — состояние скрытого нейрона. Найденный в процессе обучения вектор ξ_i затем используется для генерации новых данных, подчиняющихся совместному распределению σ и h:

$$P(\sigma; h) \propto \exp\left[-\beta \frac{E(\sigma; h)}{\sqrt{N}}\right]. \tag{2.2}$$

Параметр β, обратный температуре (1.1), описывает важность каждого выявленного признака. Распределение σ может быть получено с помощью теоремы Байеса маргинализацией h из совместного распределения (2.1), (2.2):

$$P(\sigma|\,\xi) = \frac{\operatorname{ch}\left(\dfrac{\beta}{\sqrt{N}}\xi^T\sigma\right)}{\displaystyle\sum_{\sigma}\operatorname{ch}\left(\frac{\beta}{\sqrt{N}}\xi^T\sigma\right)}. \tag{2.3}$$

Нормировочная часть может быть представлена в независимом от ξ виде:

$$\sum_{\sigma}\operatorname{ch}\left(\frac{\beta}{\sqrt{N}}\xi^T\sigma\right) = \left[2\operatorname{ch}\left(\frac{\beta}{\sqrt{N}}\right)\right]^N. \tag{2.4}$$

Если обучающая выборка состоит из M независимых примеров $\{\sigma_a\}_{a=1}^M$, то для извлечения апостериорного распределения получим

$$P(\xi|\,\{\sigma^a\}) = \frac{\displaystyle\prod_a P(\{\sigma^a\}|\,\xi)}{\displaystyle\sum_{\xi}\prod_a P(\{\sigma^a\}|\,\xi)} = \frac{1}{Z}\prod_a \operatorname{ch}\frac{\beta}{\sqrt{N}}\xi^T\sigma^a, \tag{2.5}$$

где Z — статистическая сумма рассматриваемой системы; a — индекс итераций по всем примерам; T — операция транспонирования. В качестве априорного распределения вектора ξ выбирается равномерное. Большое значение параметра β показывает, что извлеченный признак играет важную роль и будет найден в большом количестве примеров. Очевидно, что при $M > 1$ модель перестает быть тривиальной, так как статистическая сумма не может быть вычислена точно. Введем определение плотности данных $\alpha = M/N$, после чего маргинализуем $P(\xi|\,\{\sigma^a\})$ [19].

2.2. Алгоритм обучения. Будем решать задачу поиска максимума маргинализованного распределения (2.5) с помощью EM-алгоритма [20], состоящего из циклического повторения двух операций:

• E-операция вычисляет ожидаемое значение вектора скрытого слоя h из аппроксимации вектора параметров ξ на текущем шаге;

• M-операция максимизирует перекрытие $q = (1/N)\sum_i \xi_i^{\text{true}}\widehat{\xi}_i$, $\widehat{\xi}_i = \arg\max_{\xi_i} P_i(\xi_i)$ и находит следующее приближение вектора $\widehat{\xi}_i$ по текущим значениям векторов h и $\widehat{\xi}_i$.

Аналогично перекрытию в спиновых стеклах при $q = 0$ исходный вектор не несет информации о целевом векторе, а при $q = 1$ хорошо описывает его. В процессе обучения q усредняется по всем наборам истинных векторов. Для сравнения с вычислением реплик усредненное перекрытие определяется как $q = \left\langle (1/N)\sum_i \xi_i^{\text{true}} \left\langle \widehat{\xi}_i \right\rangle \right\rangle$, где внутреннее усреднение — гиббсовское, внешнее — по истинным векторам. Вычисление маргинализованной вероятности представляется сложным из-за взаимодействий, поэтому пользуемся приближением самосогласованного поля. Наложим ОМБ на фактор-граф и используем концепцию обмена сообщениями [21] (рис. 3).

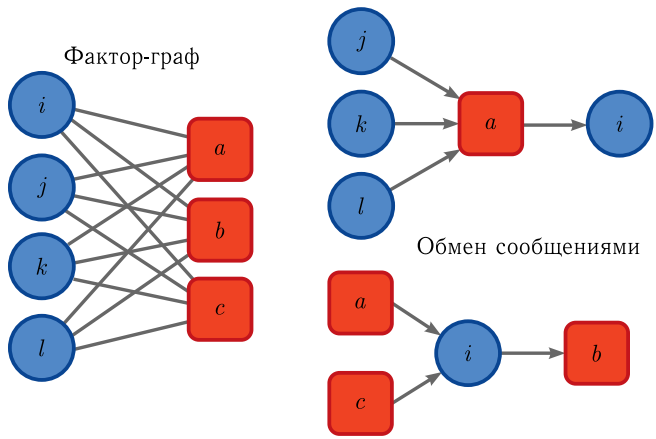

Рис. 3. Фактор-граф и представление обмена сообщениями

Уравнения передачи сообщений [22]

$$m_{i\to a} = \text{th} \sum_{b\in\partial i\setminus a} u_{b\to i}, \quad u_{b\to i} = \text{arcth}\left[\text{th}\left(\beta G_{b\to i}\right)\text{th}\left(\frac{\beta\sigma_i^b}{\sqrt{N}}\right)\right] \qquad (2.6)$$

совпадают с намагниченностью бесконечномерной модели Изинга. Здесь $G_{b\to i} = (1/N)\sum_{j\in\partial b\setminus i} \sigma_j^b m_{j\to b}$, остаточная намагниченность — $m_{j\to b} = \sum_{\xi_j} \xi_j P_{j\to b}(\xi_j)$. Интерпретация $m_{i\to a}$ — передача сообщения от i-го признака целевого вектора ξ к обучающему вектору с индексом a; σ_i^a — сила этой связи.

В приближении самосогласованного поля уравнения (2.6) будут сходиться к стационарной точке, являющейся минимумом свободной энергии [21]. Из полученного минимума m_i извлекается маргинализованная вероятность $P_i(\xi_i) = (1 + m_i\xi_i)/2$, где $m_i = \operatorname{th} \sum\limits_{b\in\partial i} u_{b\to i}$. В данном случае температура извлечения признака равна температуре, при которой было сгенерировано распределение обучающего набора данных. Каждая итерация EM-алгоритма имеет вычислительную сложность и затраты памяти порядка $O(M \cdot N)$. Для уменьшения временных затрат упростим уравнения (2.6), приняв приближение большого N. Полученное уравнение известно как уравнение Таулесса–Андерсона–Палмера:

$$Q = \frac{1}{N}\sum_i m_i^2,$$

$$G_a^{t-1} = \frac{1}{\sqrt{N}}\sum_{i\in\partial a}\sigma_i^a m_i^{t-1} - \beta\left(1 - Q^{t-1}\right)\operatorname{th}\beta G_a^{t-2}, \tag{2.7}$$

$$m_i^t \simeq \operatorname{th}\left[\sum_{b\in\partial i}\frac{\beta\sigma_i^b}{\sqrt{N}}\operatorname{th}\beta G_b^{t-1} - \frac{\beta^2 m_i^{t-1}}{N}\sum_{b\in\partial i}\left(1 - \operatorname{th}^2\beta G_b^{t-1}\right)\right].$$

Таким образом, вместо решения $2M \cdot N$ уравнений требуется решить $M + N$.

3. ОБРАБОТКА ВХОДНЫХ ДАННЫХ

3.1. Создание обучающей выборки. Создание выборки (sampling) — это гибкий способ малозатратной аппроксимации сумм и интегралов. В данной задаче алгоритм обучения требует недоступную для прямого вычисления статистическую сумму. Будем аппроксимировать ее с помощью выборки методом Монте-Карло. Идея в том, чтобы рассматривать сумму как математическое ожидание некоторого распределения и аппроксимировать его с помощью соответствующего среднего [23].

В нашем случае распределение p неизвестно; тогда воспользуемся выборкой по значимости. Важным шагом в декомпозиции слагаемого в выражении для статистической суммы

$$s = \sum_x p(x) f(x) = E_p[f(x)] \tag{3.1}$$

является решение о том, какая его часть будет выступать в роли вероятности $p(x)$, а какая — в роли случайной величины $f(x)$, математическое ожидание которой необходимо оценить. Не существует однозначно определенной декомпозиции, потому что $p(x) f(x)$ всегда можно переписать в виде

$$p(x) f(x) = g(x)\frac{p(x) f(x)}{g(x)}, \tag{3.2}$$

так что теперь выборка производится из $g(x)$, а оцениваемой величиной является pf/g.

В случае ограниченной машины Больцмана воспользоваться методом Монте-Карло напрямую невозможно, так как нельзя произвести точную выборку из распределения $p_{\text{model}}(x)$ или из выборочного распределения по значимости $g(x)$ с низкой дисперсией из-за отсутствия $p_{\text{model}}(x)$. Это происходит из-за того, что $p_{\text{model}}(x)$ представлено неориентированной графовой моделью. Тогда применим математический аппарат марковских цепей для приближенной выборки из $p_{\text{model}}(x)$.

Пусть энергетическая модель задается двумя переменными и определяет распределение $p(a, b)$. Для выбора a необходимо произвести выборку из $p(a|b)$, а для выбора b — из $p(b|a)$. Получается неразрешимая проблема «яйцо или курица». В энергетической модели порочный круг можно разорвать посредством выборки с применением марковской цепи. Основная идея — инициализировать состояние x произвольным значением. С течением времени будем случайным образом изменять x, в конечном счете x приблизится к истинной выборке из $p(x)$. Фактически марковская цепь определяется случайным состоянием x и переходным распределением $T\left(x^0 \middle| x\right)$, задающим вероятность того, что случайное изменение переведет состояние x в состояние x^0. Траектория такого процесса — многократный переход из состояния x в состояние x^0 в соответствии с распределением $T\left(x^0 \middle| x\right)$.

Для счетного множества состояний случайной величины x любое его состояние можно параметризовать — представить целым положительным числом x. Посмотрим, что произойдет, если параллельно выполнять бесконечно много марковских цепей. Все состояния марковских цепей выбираются из некоторого распределения $g^{(t)}(x)$, где t — число уже произведенных шагов. В начальный момент $g^{(0)}(x)$ — некоторое распределение, использованное для произвольной инициализации x для каждой марковской цепи. Затем на $g^{(t)}(x)$ оказывают влияние все уже произведенные шаги марковской цепи. Наша цель — добиться, чтобы $g^{(t)}(x)$ сходилось к $p(x)$. Благодаря параметризации в терминах целого положительного x можно описать распределение вероятности g с помощью вектора v, такого, что

$$g(x = i) = v_i. \tag{3.3}$$

Посмотрим, что происходит, когда состояние x одной марковской цепи изменяется на x^0. Вероятность, что новым состоянием будет x^0, равна

$$g^{(t+1)}(x') = \sum_x q^{(t)} T(x'|x). \tag{3.4}$$

Действие оператора перехода T можно представить с помощью матрицы A. Определим A следующим образом:

$$A_{ik} = T\left(x^0 = i \middle| x = k\right). \tag{3.5}$$

С помощью v и A можно описать изменение распределения всех марковских цепей, выполняемых параллельно после перехода состояния:

$$v^{(t)} = Av^{(t-1)}. \tag{3.6}$$

Перепишем формулу (3.5), воспользовавшись определением (3.6). Изменение состояния марковской цепи соответствует умножению на матрицу A. Иными словами, весь процесс можно описать как возведение матрицы A в степень:

$$v^{(t)} = A^t v^{(0)}. \tag{3.7}$$

Столбцы матрицы A представляют распределения вероятностей, такие матрицы называются *стохастическими*. Если существует ненулевая вероятность перехода из любого состояния x в любое другое состояние x^0 для некоторой степени t, то по теореме Перрона–Фробениуса наибольшее собственное значение матрицы вещественно и равно 1. Видно, что с течением времени все собственные значения возводятся в степень:

$$v^{(t)} = \left(V \operatorname{diag}(\lambda) V^{-1}\right)^{(t)} v^{(0)} = V \operatorname{diag}(\lambda)^t V^{-1} v^{(0)}. \tag{3.8}$$

В результате все собственные значения, не равные 1, стремятся к нулю. При некоторых довольно мягких условиях гарантируется, что A имеет только один собственный вектор с собственным значением 1. Поэтому процесс сходится к стационарному распределению, которое иногда называют *равновесным распределением*.

Условие для каждого шага в пределе выглядит как

$$v' = Av = v. \tag{3.9}$$

Выражение (3.9) не что иное, как уравнение собственного вектора. Чтобы оказаться стационарной точкой, v должен быть собственным вектором с собственным значением 1. Условие гарантирует, что после достижения стационарного распределения последующее применение процедуры переходной выборки не изменяет распределения состояний всех марковских цепей. После достижения марковской цепью равновесного распределения из него можно выбирать бесконечно много примеров. Все они имеют одинаковое распределение, но любые два соседних примера сильно коррелированы между собой. Поэтому конечная последовательность примеров может оказаться недостаточно репрезентативной выборкой из равновесного распределения. Один из путей смягчения этой проблемы — возвращать каждый n-й пример, чтобы оценка статистики была в меньшей степени смещена из-за корреляции между ближайшими соседями.

Еще одна трудность связана с тем, что заранее неизвестно, сколько нужно выполнить шагов для достижения равновесного распределения. Этот промежуток времени иногда называют временем перемешивания. Проверить, достигла ли марковская цепь равновесия, тоже трудно. В теории утверждается, что цепь сходится, но не более того. Анализ марков-

ской цепи по воздействию матрицы A на вектор вероятностей v показывает, что цепь достигла равновесия, когда A^t потеряла практически все собственные значения A, кроме единственного, равного 1. Это означает, что абсолютная величина второго по величине собственного значения определяет время перемешивания. На практике невозможно представить марковскую цепь матрицей, так как число возможных состояний вероятностной модели экспоненциально зависит от числа переменных. Поэтому представить v, A или собственные значения A не получится. Из-за этого и других препятствий неизвестно, достигнуто ли равновесие. Вместо этого можно дать цепи проработать какое-то время, которое будем считать достаточным, исходя из грубой оценки, и применим эвристические методы, например бутстрэп-анализ, чтобы понять, перемешалась ли цепь.

3.2. Выборка по Гиббсу и темперирование. Как убедиться, что $g(x)$ — нужное распределение? Одно из решений — вывести T из заданного обученного распределения p_{model}. Простой и эффективный способ построения марковской цепи, которая производит выборку из p_{model}, дает выборка по Гиббсу, когда выборка из $T\left(x^0 \middle| x\right)$ производится путем выбора одной величины x_i и выборки ее значений из p_{model} при условии рассмотрения соседей в неориентированном графе, определяющем структуру энергетической модели. Можно одновременно производить выборку нескольких величин, если только они условно независимы при условии рассмотрения всех своих соседей. Так, для ОМБ из всех скрытых блоков можно производить выборку одновременно, потому что они независимы друг от друга при условии рассмотрения всех видимых блоков. Аналогично можно одновременно производить выборку из всех видимых блоков в силу условной независимости их друг от друга при условии рассмотрения всех скрытых блоков. Если одновременно обновляется несколько величин, то говорят о блочной выборке по Гиббсу.

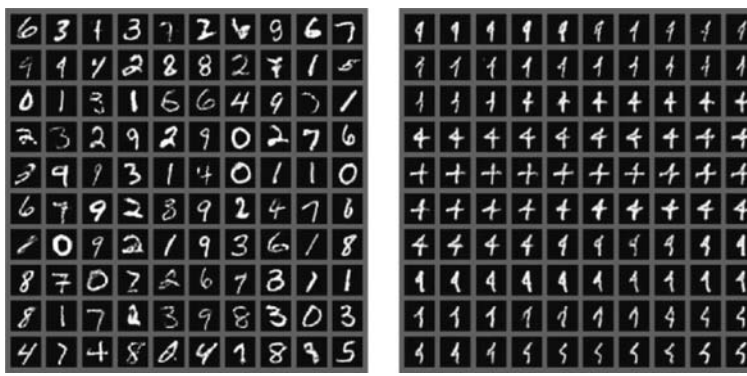

Рис. 4. Пример медленного перемешивания в глубоких моделях

Главная проблема методов Монте-Карло — плохое перемешивание — проиллюстрирована рис. 4. Можно считать, что методы Монте-Карло непреднамеренно выполняют подобие зашумленного градиентного спуска вдоль функции энергии или зашумленного восхождения на вершину функции вероятности относительно состояния цепи. Цепь демонстрирует тенденцию к малым шагам в пространстве состояний марковской цепи из конфигурации $x^{(t-1)}$ в конфигурацию $x^{(t)}$.

Предпочтение отдается переходам в конфигурации с более низкой энергией. Начав с маловероятной конфигурации с энергией выше, чем для типичных примеров из $p(x)$, цепь стремится постепенно уменьшать энергию состояния и лишь изредка переходит на другую моду. Примеры идут слева направо сверху вниз. Слева — ближние соседи из выборки по Гиббсу, примененной к глубокой машине Больцмана, обученной на наборе данных MNIST (Modified National Institute of Standards and Technology — объемная база данных образцов рукописного написания цифр). Соседние примеры похожи друг на друга. Поскольку выборка по Гиббсу производится в глубокой графической модели, это сходство основано, скорее, на семантике, чем на визуальных признаках, но все равно марковской цепи трудно перейти из одной моды распределения в другую, например путем изменения цифры. Справа — ближние соседи выборки из порождающей состязательной сети. Поскольку модель является ориентированной и порождает примеры независимо, то проблемы перемешивания нет [23].

После того как цепь нашла область низкой энергии, которую называем модой, она начинает перемещаться вокруг этой моды, совершая своего рода случайное блуждание. Время от времени цепь покидает эту моду и либо возвращается к ней, либо, если найдет путь выхода, переходит к другой моде. Проблема в том, что для многих распределений пути выхода встречаются редко, поэтому марковская цепь продолжает выбирать примеры из одной и той же моды слишком долго.

Вероятность перехода из одной моды в соседнюю определяется формой энергетического барьера между модами. Переходы между модами, разделенными высоким барьером (областью низкой вероятности), экспоненциально менее вероятны. Несколько приемов повышения скорости перемешивания основаны на построении альтернативных вариантов целевого распределения, в котором пики не такие высокие, а окружающие их долины не такие низкие. В энергетических моделях сделать это особенно удобно с помощью параметра β, контролирующего остроту пика. Когда температура стремится к нулю, а β устремляется к бесконечности, энергетическая модель становится детерминированной. Если же температура стремится к бесконечности, а β — к нулю, то распределение становится равномерным. Обычно модель обучают при $\beta = 1$, но можно использовать и другие температуры, в частности $\beta < 1$.

Темперирование (*tempering*) — это общая стратегия быстрого перемешивания мод путем выборки примеров с $\beta < 1$. Марковские цепи, ос-

нованные на темперированных переходах, временно производят выборку из высокотемпературных распределений, чтобы перемешать разные моды, а затем возобновляют выборку из распределения с единичной температурой. Другой подход — использование параллельного темперирования, когда марковская цепь параллельно имитирует много различных состояний при разных температурах. Высокотемпературные состояния медленно перемешиваются, а низкотемпературные (с температурой 1) дают верные выборки из модели. Оператор перехода включает стохастический обмен состояний из двух разных температурных режимов так, чтобы пример с достаточно большой вероятностью из высокотемпературного состояния мог перескочить в низкотемпературное.

Оба подхода темперирования часто применяются для ОМБ. Тем не менее в настоящее время они не позволили далеко продвинуться в решении проблемы выборки из сложных энергетических моделей. Возможно, дело в том, что существуют критические температуры, в окрестности которых температурный переход должен быть очень медленным, и только тогда темперирование оказывается эффективным.

3.3. Измерение температуры обучающей выборки. Можно ли вывести температуру, использованную для создания самих данных? Если да, то изучим типичные свойства фазовых переходов, присущих исследуемой системе. Это возможно сделать, применив байесовское правило еще раз [19]. Апостериорная вероятность β, задаваемая входными данными $\{\sigma^a\}_{a=1}^M$, описывается формулой

$$P\left(\beta|\,\{\sigma^a\}\right) = \sum_{\xi} P\left(\beta,\xi|\,\{\sigma^a\}\right) = \frac{P\left(\{\sigma^a\}|\,\xi,\beta\right) P_0\left(\xi,\beta\right)}{\int d\beta \sum\limits_{\xi} P\left(\{\sigma^a\}|\,\xi,\beta\right) P_0\left(\xi,\beta\right)} =$$

$$= \frac{1}{Z\left(\{\sigma^a\}\right)} \sum_{\xi} \exp\left\{-NM \ln\left[2\,\mathrm{ch}\left(\frac{\beta}{\sqrt{N}}\right)\right]\right\} \prod_a \mathrm{ch}\left(\frac{\beta}{\sqrt{N}} \xi^T \sigma^a\right) \propto$$

$$\propto \exp\left(-M\frac{\beta^2}{2}\right) Z\left(\beta,\{\sigma^a\}\right), \quad (3.10)$$

где P_0 — равномерное априорное распределение гиперпараметров. Максимизируя апостериорную вероятность, получим, что самосогласованное уравнение для β должно удовлетворять условию

$$\frac{\partial \ln Z\left(\beta|\,\{\sigma^a\}\right)}{\partial\beta} = N\alpha\beta. \quad (3.11)$$

В левой части выражения (3.11) находится энергия $(-N\varepsilon)$, описываемая уравнением передачи сообщений (2.6). Для случая малых N выражение для β выглядит как $\beta = \sqrt{N}\,\mathrm{arcth}\left(-\varepsilon/\alpha\sqrt{N}\right)$, откуда видно происхождение (3.11) для случая больших N, где ε — энергия, при-

ходящаяся на один нейрон, вычисляемая по формуле $N\varepsilon = -\sum_i \Delta\varepsilon_i + + (N-1)\sum_a \Delta\varepsilon_a$, в которой $\Delta\varepsilon_i$ и $\Delta\varepsilon_a$ выражаются как

$$\Delta\varepsilon_i = \frac{\sum\limits_{a\in\partial i} \mathcal{H}_{a\to i}(+1) + \prod\limits_{a\in\partial i} \mathcal{G}_{a\to i} \sum\limits_{a\in\partial i} \mathcal{H}_{a\to i}(-1)}{\beta + \beta \prod\limits_{a\in\partial i} \mathcal{G}_{a\to i}}, \qquad (3.12)$$

$$\Delta\varepsilon_a = \beta\Xi_a^2 + G_a \mathrm{th}\,(\beta G_a)\,,$$

с использованием обозначений

$$\mathcal{G}_{a\to i} = \mathrm{e}^{-2u_{a\to i}},\ \ \Xi_{a\to i}^2 \simeq \frac{1}{N}\sum_{j\in\partial a\setminus i}\left(1-m_{j\to a}^2\right),\ \ \Xi_a^2 = \frac{1}{N}\sum_{i\in\partial a}\left(1-m_{i\to a}^2\right),$$

$$\mathcal{H}_{a\to i} = \beta^2\Xi_{a\to i}^2 + \beta\left(G_{a\to i} + \frac{\sigma_i^a\xi_i}{\sqrt{N}}\right)\mathrm{th}\,\beta\left(G_{a\to i} + \frac{\sigma_i^a\xi_i}{\sqrt{N}}\right).$$

Начиная с некоторого начального значения β, можно итеративно подбирать его для достижения сходимости с заданной точностью.

Температура признака связана с его значимостью в наборе данных. Как только известна температура, можно узнать, насколько явно виден скрытый признак в наборе данных, и определить критический размер обучающей выборки для обучения без учителя. Более того, можно исследовать критические явления, такие как нарушение симметрии реплик.

В заключение теоретической части отметим, что было проведено аналитическое исследование ограниченной машины Больцмана методами статистической физики и теории вероятностей, обоснован переход от спиновых стекол к семейству нейронных сетей, описаны методы численного приближения статистической суммы модели и получено уравнение для извлечения гиперпараметра алгоритма обучения.

4. МОДЕЛИРОВАНИЕ

Исследуем поведение системы с помощью описанных методов. Рассмотрим изменение энтропии, приходящейся на один нейрон, в зависимости от изменения размера обучающей выборки при фиксированном размере сети N (рис. 5). Из графика видно, что при определенной плотности данных наступает энтропийный кризис — энтропия становится отрицательной, что запрещено для систем с дискретными степенями свободы. Энтропийный кризис отделяет экспоненциальный режим изменения q от подэкспоненциального. Переход происходит в пределах экспоненциального режима, если важность признака достаточно велика. На рис. 5 изображены результаты, полученные в приближении самосогласованного поля моделированием алгоритма передачи простых сообщений (ППС) в предположении симметрии реплик. Погрешность метода меньше

обозначений экспериментальных точек. Экспериментальные точки могут быть приближены прямой для каждого случая, что подтверждает согласование эксперимента с теорией. Энтропия показывает, как соотносится изменение размера выборки с размером сети N. Оценим работу двух алгоритмов вычисления перекрытия с помощью уравнений (2.6) и (2.7). На рис. 6 приведен график сходимости перекрытия, полученного вычислением аппроксимированных уравнений передачи сообщений (АПС) (2.7)

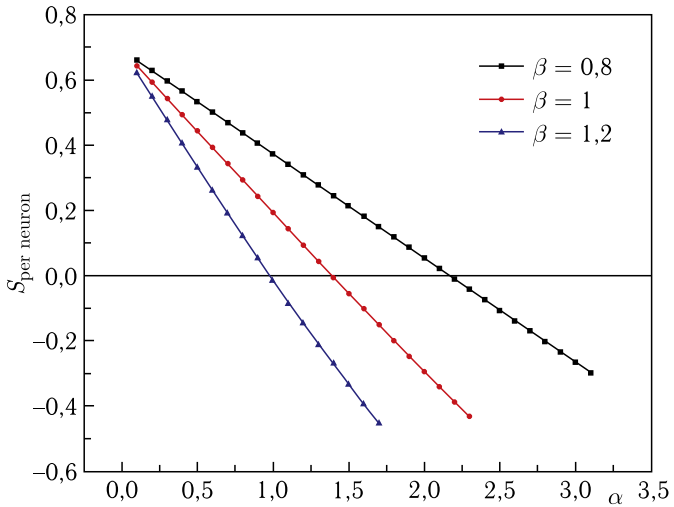

Рис. 5. Зависимость энтропии, приходящейся на нейрон, от плотности данных α

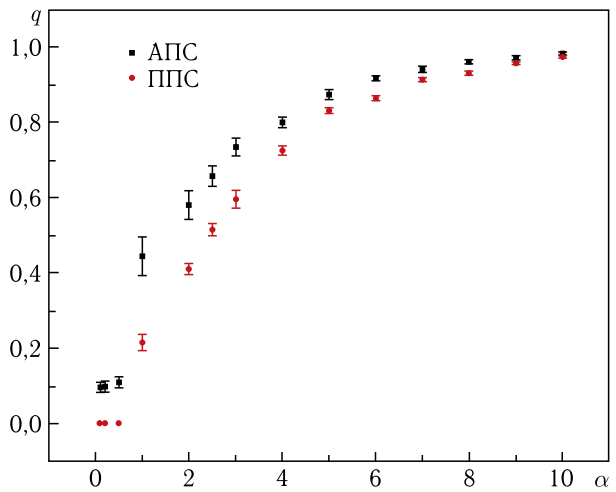

Рис. 6. Сходимость перекрытия в зависимости от плотности данных α

к перекрытию, полученному решением уравнения простой передачи сообщений (2.6) в зависимости от плотности данных α.

В результате сравнения можно сделать вывод о возможности использования обоих алгоритмов без видимой потери качества при больших α. Как было сказано в теоретической части, второй алгоритм имеет аддитивную временную сложность в зависимости от количества нейронов N и количества обучающих примеров M, в то время как первый — мультипликативную. Для выбора температуры извлечения признаков обучающей выборки построим график (рис. 7).

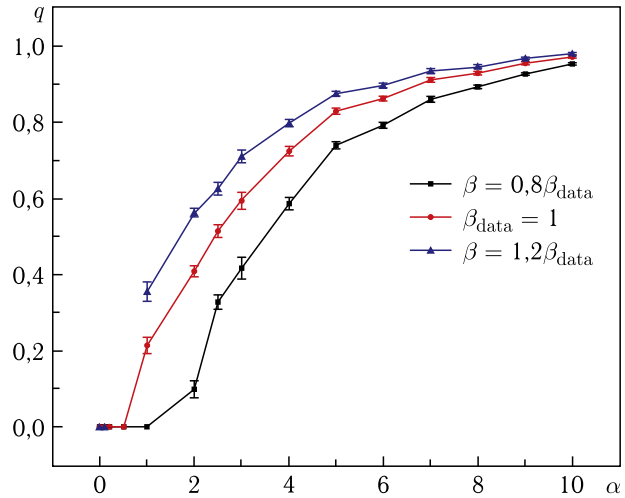

Рис. 7. Изменение скорости расчета от плотности данных α

Очевидно, что оптимальная температура обработки равна температуре создания данных. В таком случае полезным является наибольшее число обучающих примеров из выборки. В аспекте малой плотности данных — количество обучающих примеров не всегда может достигать и превышать размеры нейронной сети — важно найти температуру обучающей выборки.

На рис. 7 приведено изменение скорости расчета перекрытия при изменении температуры обработки данных относительно температуры создания данных в зависимости от плотности данных α. При температуре обработки, меньшей температуры выборки, перекрытие значительно возрастает при небольших α, но при совсем малых α (порядка 0,2–0,5) сходимость алгоритма не наступает.

Зафиксируем силу встроенного признака — $\beta/\sqrt{N}$ на отметке 0,1 — и изучим влияние размера сети на сохранение одинаковой силы признаков. На рис. 8 приведено изменение скорости расчета перекрытия при варьировании размеров сети с сохранением силы встроенного признака

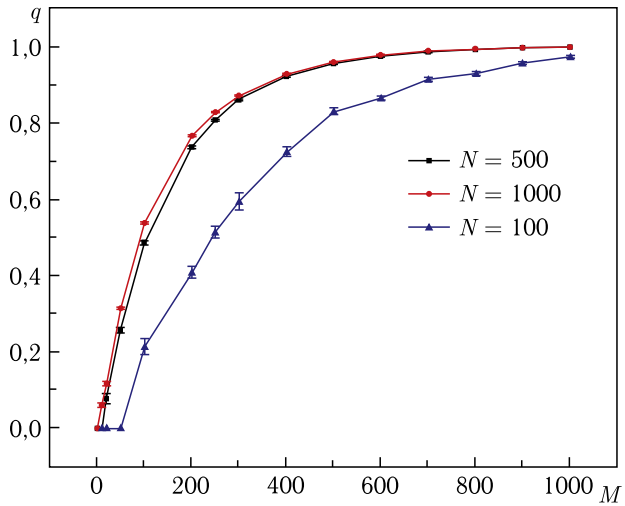

Рис. 8. Изменение скорости расчета перекрытия в зависимости от размера обучающей выборки M

$\beta/\sqrt{N}$ на отметке 0,1 в зависимости от размера обучающей выборки M. При заданном количестве примеров сеть с бóльшим размером дает лучший результат перекрытия. Однако график перекрытия (см. рис. 8) достигает насыщения, когда $N \approx 1000$ при относительно большом размере обучающей выборки M. Решение задачи вариации размера сети позволяет получить оптимальное значение еще одного гиперпараметра алгоритма обучения. Рассмотрим работу EM-алгоритма для извлечения истинной температуры обучающего набора. Стартуя с заданного начального приближения, алгоритм итеративно подбирает значение β. При этом (согласно описанной теории передачи сообщений) на E-шаге обновляются узлы, содержащие сообщения, на M-шаге — обратная температура. Для сглаживания численного решения используется линейная комбинация $\beta(t) = \eta\beta(t) + (1-\eta)\beta(t-1)$, где t — временной шаг, $\eta \in [0,1]$ — скорость обучения (демпинг-фактор).

На рис. 9 демонстрируется сходимость алгоритма к истинному значению β_{true} при различных объемах обучающей выборки. Из графика зависимости подбираемой температуры видно (см. рис. 9), что для различных размеров обучающей выборки результат работы алгоритма различается. При построении траектории движения алгоритма была взята одна из 10 тестовых, успешно достигшая сходимости. В силу произвольной инициализации весов модели равномерным распределением полученная температура для каждой из траекторий отличается от других.

Очевидно, что существует оптимальное значение M, при котором $|\beta_{\text{true}} - \beta| \to 0$. Для нахождения указанного минимума воспользуемся

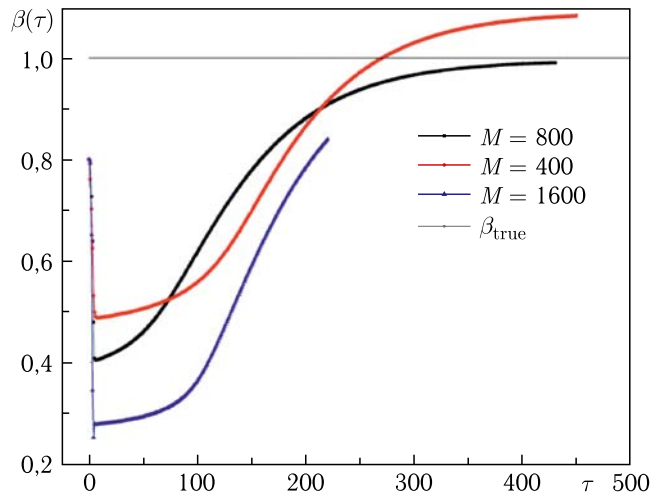

Рис. 9. Сходимость алгоритма при различных объемах обучающей выборки

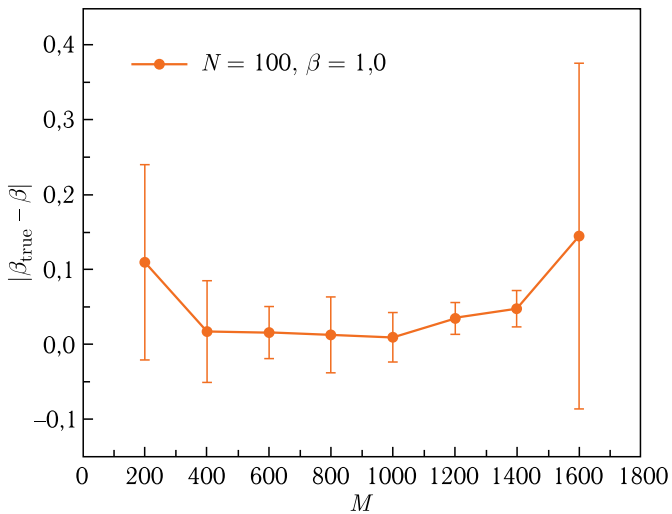

Рис. 10. Нахождение оптимального размера обучающей выборки

поиском по сетке $[100, 1600]$ с шагом 200. Для оценки $|\beta_{\text{true}} - \beta|$ будем использовать усредненное по 10 сошедшимся траекториям для каждой итерации при данном M. В качестве метрики ошибки алгоритма возьмем среднеквадратичное отклонение. На рис. 10 показано нахождение оптимального размера обучающей выборки M для извлечения температуры

создания данных β_{true} поиском по равномерной сетке [200, 1600] с шагом 200.

Для сгенерированных при $\beta_{\text{true}} = 1$ данных по мере увеличения размера выборки подбираемое значение β приближается к истинному значению. Как показано на рис. 10, зависящее от времени (шага итерации) выводимое значение сначала падает до более низкого, а затем постепенно приближается к истинному значению. Причем чем больше величина обучающих данных M, тем сильнее начальное падение β.

ЗАКЛЮЧЕНИЕ

В работе рассмотрены методы статистической физики для анализа спиновых стекол. Предложенные методы позволяют произвести эквивалентный переход к семейству нейронных сетей. В рассматриваемых схемах используют байесовский вывод на скрытых марковских цепях, в отличие от большинства известных предложений, основывающихся на градиентном спуске. Преимуществом предложенного подхода является возможность аналитического рассмотрения системы. В заключительной части работы описаны методы извлечения гиперпараметров алгоритма работы ограниченной машины Больцмана. Показано, какие наборы значений могут привести к построению максимально эффективной процедуры обучения извлечению скрытых признаков в неразмеченных данных.

В качестве основных результатов можно отметить следующие.

• Приведен алгоритм для приближения расчета статистической суммы системы, состоящей из M реплик.

• Экспериментально продемонстрирована возможность уменьшения вычислительной сложности алгоритма передачи сообщений при помощи аппроксимации уравнений.

• Исследованы критические явления в ограниченной машине Больцмана — энтропийный кризис, различие в температурах создания и обработки обучающей выборки.

• Найдены оптимальные параметры синтетической выборки для нейронной сети с $N = 100$ видимых нейронов и $\beta = 1{,}0$.

Полученные результаты могут позволить существенно ускорить алгоритм обучения ограниченной машины Больцмана, избежать энтропийного кризиса (расхождения работы алгоритма) и исследовать предоставленный набор обучающих данных на репрезентативность.

Благодарности. Работа выполнена при поддержке междисциплинарной научно-образовательной школы Московского государственного университета им. М. В. Ломоносова «Фотонные и квантовые технологии. Цифровая медицина».

СПИСОК ЛИТЕРАТУРЫ

1. *Samuel A. L.* Some Studies in Machine Learning Using the Game of Checkers // IBM J. Res. Develop. 1959. V. 3, No. 3. P. 210–229.

2. *Иноземцева Н. Г., Перепёлкин Е. Е., Садовников Б. И.* Оптимизация алгоритмов задач математической физики для графических процессоров. М.: Изд-во Моск. ун-та, 2012. 256 с.
3. *Перепёлкин Е. Е., Садовников Б. И., Иноземцева Н. Г.* Вычисления на графических процессорах (GPU) в задачах математической и теоретической физики. Сер. «Классический учеб. МГУ». М.: URSS, 2019. 240 с.
4. *Mezard M., Parisi G., Virasoro M. A.* Spin Glass Theory and Beyond. World Sci. Publ., 1987. P. 317.
5. *Sherrington D.* Neural Networks: The Spin Glass Approach // Math. Approaches to Neural Networks. 1993. V. 51. P. 261–291.
6. *Lake B. M., Salakhutdinov R., Tenenbaum J. B.* Human-Level Concept Learning through Probabilistic Program Induction // Science. 2015. V. 350. P. 1332–1338.
7. *Кадурин А. А., Николенко С. И., Архангельская Е. В.* Глубокое обучение. Погружение в мир нейронных сетей. СПб.: Питер, 2018.
8. *Hinton G.* A Practical Guide to Training Restricted Boltzmann Machines. UTML TR 2010–003. http://www.cs.toronto.edu/ hinton/absps/guideTR.pdf.
9. *Hinton G. E., Osindero S., Teh Y.-W.* A Fast Learning Algorithm for Deep Belief Nets // Neural Comput. 2006. Part. 18, No. 7. P. 1527–1554.
10. *Agliari E., Barra A., Tirozzi B.* Free Energies of Boltzmann Machines: Self-Averaging, Annealed and Replica Symmetric Approximations in the Thermodynamic Limit. arXiv:1810.11075.
11. *Song Juyong, Marsili M., Jo Junghyo.* Resolution and Relevance Trade-offs in Deep Learning. arXiv:1710.11324 [cs.LG].
12. *Salakhutdinov R., Mnih A., Hinton G.* Restricted Boltzmann Machines for Collaborative Filtering // Proc. of the 24th Intern. Conf. on Machine Learning. ACM. 2007.
13. *Barra et al.* On the Equivalence among Hopfield Neural Networks and Restricted Boltzmann Machines // Neural Netw. 2012. V. 34. P. 1–9.
14. *Боголюбов Н. Н.* Собр. науч. тр.: В 12 т. Т. 6. Равновесная статистическая механика 1945–1986. М.: Наука, 2006. 520 с.
15. *Castellani T., Cavagna A.* Spin-Glass Theory for Pedestrians. arXiv:cond-mat/0505032[cond-mat.dis-nn].
16. *Nishimori H.* Statistical Physics of Spin Glasses and Information Processing: An Introduction. Oxford: Oxford Univ. Press, 2001.
17. *Parisi G.* // Phys. Lett. A. 1979. V. 73. P. 203–205.
18. *Parisi G.* // J. Phys. A. 1980. V. 13. P. L115–L121, 1101–1112, and 1887–1895.
19. *Smolensky P.* Information Processing in Dynamical Systems: Foundations of Harmony Theory. Ch. 6. 1986.
20. *Hartnetta G. S., Parker E., Geist E.* Replica Symmetry Breaking in Bipartite Spin Glasses and Neural Networks. arXiv:1803.06442[cond-mat.dis-nn].
21. *Bengio Y., Courville A., Vincent P.* Representation Learning: A Review and New Perspectives. Pattern Analysis and Machine Intelligence // IEEE Trans. 2013. V. 35. P. 1798–1828.
22. *Haiping Huang.* Statistical Mechanics of Unsupervised Feature Learning in a Restricted Boltzmann Machine with Binary Synapses. RIKEN Brain Science Inst., Wako-shi, Saitama. 2018.

23. *Воронцов К. В.* Математические методы обучения по прецедентам (теория обучения машин). С. 32–40; http://www.machinelearning.ru/wiki/index.php?title=Mo.
24. *Mezard M., Montanari A.* Information, Physics, and Computation. Oxford: Oxford Univ. Press, 2009.
25. *Haiping Huang, Taro Toyoizumi.* Unsupervised Feature Learning from Finite Data by Message Passing: Discontinuous versus Continuous Phase Transition // Phys. Rev. E. 2016. V. 94. P. 062310.
26. *Гудфеллоу Я., Бенджио И., Курвилль А.* Глубокое обучение. М.: ДМК Пресс, 2018. С. 496–507.