

© 2018 г. О.Ю. БАХТЕЕВ ([bakhteev@phystech.edu](mailto:bakhteev@phystech.edu))  
(Московский физико-технический институт),  
В.В. СТРИЖОВ, д-р физ.-мат. наук ([strijov@phystech.edu](mailto:strijov@phystech.edu))  
(Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН)

## ВЫБОР МОДЕЛЕЙ ГЛУБОКОГО ОБУЧЕНИЯ СУБОПТИМАЛЬНОЙ СЛОЖНОСТИ<sup>1</sup>

Рассматривается задача выбора моделей глубокого обучения субоптимальной сложности. Под сложностью модели понимается минимальная длина описания совокупности выборки и модели классификации или регрессии. Под субоптимальной сложностью понимается приближенная оценка минимальной длины описания, полученная с использованием байесовского вывода и вариационных методов. Вводятся вероятностные предположения о распределении параметров. На основе байесовского вывода предлагается функция правдоподобия модели. Для получения оценки правдоподобия применяются вариационные методы с использованием градиентных алгоритмов оптимизации. Проводится вычислительный эксперимент на нескольких выборках.

*Ключевые слова:* классификация, регрессия, глубокое обучение, выбор модели, байесовский вывод, вариационный вывод, сложность.

### 1. Введение

Проблема выбора модели является одной из ключевых задач машинного обучения. Под моделью понимается суперпозиция функций, решающая задачу классификации или регрессии. В данной работе в качестве критерия выбора модели предлагается субоптимальная сложность модели. Под сложностью модели понимается *минимальная длина описания* [1], т.е. минимальное количество информации, которое требуется передать о модели и о выборке. Вычисление минимальной длины описания модели является вычислительно сложной процедурой. В работе предлагается получение ее приближенной оценки, основанной на связи минимальной длины описания и *правдоподобия модели* [1]. В общем случае данная величина является трудновычислимой. Для получения оценки правдоподобия используются вариационные методы получения оценки правдоподобия [2], основанные на аппроксимации неизвестного другим заданным распределением. Под субоптимальной сложностью понимается вариационная оценка правдоподобия модели.

Одним из методов получения приближенного значения правдоподобия модели является вариационный метод получения нижней оценки правдоподобия [2]. В [3] рассматривается алгоритм получения вариационной нижней оценки правдоподобия для оптимизации гиперпараметров моделей глубокого обучения. В [4] рассматривается стохастический градиентный спуск в качестве оператора, порождающего распределение, аппроксимирующее апостериорное распределение параметров модели. В

---

<sup>1</sup>Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект 16-37-00488) и РФ (соглашение № 05.Y09.21.0018).

работе отмечается, что стохастический градиентный спуск не оптимизирует вариационную оценку правдоподобия, а приближает ее только до некоторого числа итераций оптимизации.

Одна из проблем построения моделей глубокого обучения — большое число параметров модели [5], которое достигает нескольких миллионов, а оптимизация модели достигает десятков дней [6]. Задача выбора модели глубокого обучения включает в себя выбор стратегии построения модели, эффективной по вычислительным ресурсам.

В данной работе предлагается метод получения вариационной нижней оценки правдоподобия модели с использованием модифицированного алгоритма стохастического градиентного спуска. Модификация заключается в добавлении шумовой компоненты. Эта компонента позволяет получить более точные оценки правдоподобия модели для сравнения моделей и выбора наиболее адекватной из них. Рассматривается ряд модификаций базового алгоритма. В качестве базового алгоритма выступает алгоритм оптимизации параметров модели с использованием стохастического градиентного спуска без контроля переобучения. Он заключается в итеративном вычислении градиента по параметрам от функции правдоподобия обучающей выборки и изменении значений параметров с его учетом. Приводится сравнение с алгоритмом получения вариационной нижней оценки, представленном в [3]. Рассматриваются следующие модификации базового алгоритма: оптимизация с кросс-валидацией с использованием и без использования регуляризации модели, алгоритм получения вариационной оценки правдоподобия модели с применением нормального распределения, алгоритм получения вариационной оценки правдоподобия с использованием стохастического градиентного спуска, алгоритм получения вариационной оценки правдоподобия с использованием стохастической динамики Ланжевена. Данные алгоритмы решают следующие проблемы оптимизации моделей градиентным спуском: оптимизация модели с меньшими затратами вычислительных ресурсов, быстрая сходимость оптимизации, контроль переобучения и выбор наиболее адекватной модели. Под переобучением понимается потеря обобщающей способности модели с увеличением правдоподобия обучающей выборки [7]. Переобучение характерно для моделей с большим количеством параметров, сопоставимым с мощностью обучающей выборки, что встречается в случае выбора моделей глубокого обучения [5, 6]. Также алгоритмы имеют дальнейшую возможность применения к градиентным алгоритмам оптимизации гиперпараметров, описанным в [8].

Свойства представленных в данной работе алгоритмов исследуются на выборках, на которых проверялась работа алгоритма вероятностного обратного распространения ошибок [9], где авторы акцентируют внимание на оптимизации параметров модели.

## 2. Постановка задачи оптимизации правдоподобия модели

Задана выборка

$$(1) \quad \mathcal{D} = \{(\mathbf{x}_i, y_i)\}, i = 1, \dots, m,$$

состоящая из множества пар «объект-метка»

$$\mathbf{x}_i \in \mathbf{X} \subset \mathbb{R}^n, \quad y_i \in \mathbf{y} \subset \mathbb{Y}.$$

Метка  $y$  объекта  $\mathbf{x}$  принадлежит либо множеству:  $y \in \mathbb{Y} = \{1, \dots, Z\}$  в случае задачи классификации, где  $Z$  — число классов, либо некоторому подмножеству вещественных чисел  $y \in \mathbb{Y} \subseteq \mathbb{R}$  в случае задачи регрессии.

Моделью глубокого обучения  $\mathbf{f}$  назовем суперпозицию функций

$$(2) \quad \mathbf{f}(\mathbf{w}, \mathbf{X}) = \mathbf{f}_1(\mathbf{f}_2(\dots \mathbf{f}_K(\mathbf{w}, \mathbf{X}))) : \mathbb{R}^{m \times n} \rightarrow \mathbb{Y}^m,$$

где  $\mathbf{f}_k$  — подмодели, параметрическое семейство дважды дифференцируемых по параметрам вектор-функций,  $k \in \{1, \dots, K\}$ ;  $\mathbf{w} \in \mathbb{R}^u$  — вектор параметров моделей. Для каждой модели определена функция правдоподобия  $p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{f})$ , где  $\mathbf{x}$  — строка матрицы  $\mathbf{X}$ ,  $\mathbf{y}$  — вектор меток зависимой переменной  $y$ . Множество всех рассматриваемых моделей обозначим  $\mathfrak{F}$ . Для каждой модели  $\mathbf{f}$  из конечного множества моделей  $\mathfrak{F}$  задано априорное распределение параметров  $p(\mathbf{w}|\mathbf{f})$ .

*Определение 1.* Сложностью модели  $\mathbf{f}$  назовем правдоподобие модели:

$$(3) \quad p(\mathbf{y}|\mathbf{X}, \mathbf{f}) = \int_{\mathbf{w} \in \mathbb{R}^u} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{f})p(\mathbf{w}|\mathbf{f})d\mathbf{w}.$$

Модели  $\mathbf{f} \in \mathfrak{F}$  имеют различные размерности  $u$  соответствующих векторов параметров. Также заданы различные априорные распределения их параметров  $p(\mathbf{w}|\mathbf{f})$ .

*Определение 2.* Модель классификации  $\mathbf{f}$  назовем оптимальной среди моделей  $\mathfrak{F}$ , если достигается максимум интеграла (3).

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

$$(4) \quad p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{f}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{f})p(\mathbf{w}|\mathbf{f})}{p(\mathbf{y}|\mathbf{X}, \mathbf{f})}.$$

*Пример 1.* Рассмотрим задачу линейной регрессии:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}),$$

где  $\mathbf{A}$  — диагональная матрица. Правдоподобие зависимой переменной имеет вид

$$(5) \quad p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{f}) = (2\pi)^{-\frac{m}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w})\right),$$

априорное распределение параметров модели имеет вид

$$(6) \quad p(\mathbf{w}|\mathbf{f}) = (2\pi)^{-\frac{n}{2}} |\mathbf{A}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{w}^\top \mathbf{A} \mathbf{w}\right).$$

Правдоподобие модели (3) в этом примере вычисляется аналитически [10]:

$$(7) \quad p(\mathbf{y}|\mathbf{X}, \mathbf{f}) = (2\pi)^{-\frac{m}{2}} |\mathbf{A}|^{\frac{1}{2}} |\mathbf{H}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^\top(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})\right) \exp\left(-\frac{1}{2}\hat{\mathbf{w}}^\top \mathbf{A} \hat{\mathbf{w}}\right),$$

где  $\hat{\mathbf{w}}$  — значение наиболее вероятных (4) параметров модели:

$$\hat{\mathbf{w}} = \arg \max p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{f}) = (\mathbf{A} + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

$\mathbf{H}$  — гессиан функции потерь  $L$  модели:

$$\mathbf{H} = \nabla \nabla_{\mathbf{w}} \left( \frac{1}{2} (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{1}{2} \mathbf{w}^\top \mathbf{A} \mathbf{w} \right) = \mathbf{A} + \mathbf{X}^\top \mathbf{X},$$

$$L = -\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{f}).$$

В качестве функции, приближающей логарифм интеграла (3), будем рассматривать его нижнюю оценку, полученную при помощи неравенства Йенсена [2]. Получим нижнюю оценку логарифма правдоподобия модели, используя неравенство

$$\begin{aligned} (8) \quad \log p(\mathbf{y}|\mathbf{X}, \mathbf{f}) &= \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{f})}{q(\mathbf{w})} d\mathbf{w} + D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{f})) \geqslant \\ &\geqslant \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{f})}{q(\mathbf{w})} d\mathbf{w} = \\ &= -D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) + \int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{f}) d\mathbf{w}, \end{aligned}$$

где  $D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f}))$  — расстояние Кульбака–Лейблера между двумя распределениями:

$$\begin{aligned} D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) &= - \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w}, \\ p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{f}) &= p(\mathbf{y}|\mathbf{X}, \mathbf{f})p(\mathbf{w}|\mathbf{f}). \end{aligned}$$

*Определение 3.* Вариационной оценкой логарифма правдоподобия модели (3)  $\log p(\mathbf{y}|\mathbf{X}, \mathbf{f})$  называется оценка  $\log \hat{p}(\mathbf{y}|\mathbf{X}, \mathbf{f})$ , полученная аппроксимацией неизвестного апостериорного распределения  $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{f})$  заданным распределением  $q(\mathbf{w})$ .

Будем рассматривать задачу нахождения вариационной оценки как задачу оптимизации. Пусть задано множество распределений  $Q = \{q(\mathbf{w})\}$ . Сведем задачу нахождения наиболее близкой вариационной нижней оценки интеграла (3) к оптимизации вида

$$\hat{q}(\mathbf{w}) = \arg \max_{q \in Q} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{f})}{q(\mathbf{w})} d\mathbf{w}.$$

В данной работе в качестве множества  $Q$  рассматривается нормальное распределение и распределение параметров, неявно получаемое оптимизацией градиентными методами.

Оценка (8) является нижней, поэтому может давать некорректные оценки для правдоподобия (3). Для того, чтобы оценить величину этой ошибки, докажем следующее утверждение.

*Утверждение 1.* Пусть задано множество  $Q = \{q(\mathbf{w})\}$  непрерывных распределений. Максимизация вариационной нижней оценки

$$\int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w} | \mathbf{X}, \mathbf{f})}{q(\mathbf{w})} d\mathbf{w}$$

логарифма интеграла (3) эквивалентна минимизации расстояния Кульбака–Лейблера между распределением  $q(\mathbf{w}) \in Q$  и апостериорным распределением параметров  $p(\mathbf{w} | \mathbf{y}, \mathbf{X}, \mathbf{f})$ :

$$(9) \quad \hat{q} = \arg \max_{q \in Q} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w} | \mathbf{X}, \mathbf{f})}{q(\mathbf{w})} d\mathbf{w} \Leftrightarrow \hat{q} = \arg \min_{q \in Q} D_{\text{KL}}(q(\mathbf{w}) || p(\mathbf{w} | \mathbf{y}, \mathbf{X}, \mathbf{f})),$$

$$D_{\text{KL}}(q(\mathbf{w}) || p(\mathbf{w} | \mathbf{y}, \mathbf{X}, \mathbf{f})) = \int_{\mathbf{w}} q(\mathbf{w}) \frac{q(\mathbf{w})}{p(\mathbf{w} | \mathbf{y}, \mathbf{X}, \mathbf{f})} d\mathbf{w}.$$

Таким образом, задача нахождения вариационной оценки, близкой к значению интеграла (3) сводится к поиску распределения  $\hat{q}$ , аппроксимирующего распределение  $p(\mathbf{w} | \mathbf{y}, \mathbf{X}, \mathbf{f})$  наилучшим образом. Доказательство утверждения 1 см. в Приложении.

*Определение 4.* Модель  $\mathbf{f}$  назовем субоптимальной на множестве моделей  $\mathfrak{F}$  по множеству распределений  $Q$ , если модель доставляет максимум нижней вариационной оценке интеграла (9)

$$(10) \quad \max_{q \in Q} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w} | \mathbf{X}, \mathbf{f})}{q(\mathbf{w})} d\mathbf{w}.$$

Субоптимальность модели может быть также названа вариационной оптимальностью модели или LB-оптимальностью (*Lower Bound — нижняя граница*) модели.

Вариационная оценка (8) интерпретируется как оценка сложности модели по принципу минимальной длины описания [1], где первое слагаемое определяет количество информации для описания выборки, а второе слагаемое — длину описания самой модели [3].

В данной работе решается задача выбора субоптимальной модели при различных заданных множествах  $Q$ .

### 3. Методы получения вариационной оценки правдоподобия

Ниже представлены методы получения вариационных нижних оценок (10) правдоподобия (3). В первом подразделе рассматривается метод, основанный на аппроксимации апостериорного распределения  $p(\mathbf{w} | \mathbf{y}, \mathbf{X}, \mathbf{f})$  (4) многомерным гауссовым распределением с диагональной матрицей ковариаций. В последующих разделах рассматриваются методы, основанные на различных модификациях стохастического градиентного спуска.

### 3.1. Аппроксимация нормальным распределением

В качестве множества  $Q = \{q(\mathbf{w})\}$  задано параметрическое семейство нормальных распределений с диагональными матрицами ковариаций:

$$(11) \quad q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}),$$

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

Пусть априорное распределение  $p(\mathbf{w}|\mathbf{f})$  (6) параметров модели задано как нормальное:

$$p(\mathbf{w}|\mathbf{f}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}^{-1}).$$

Тогда оптимизация (9) имеет вид

$$(12) \quad \int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{f}) d\mathbf{w} - D_{\text{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) \rightarrow \max_{\mathbf{A}_q, \boldsymbol{\mu}_q},$$

где расстояние  $D_{\text{KL}}$  между двумя гауссовыми величинами рассчитывается как

$$D_{\text{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) = \frac{1}{2} (\text{Tr}[\mathbf{A}\mathbf{A}_q^{-1}] + (\boldsymbol{\mu} - \boldsymbol{\mu}_q)^T \mathbf{A}(\boldsymbol{\mu} - \boldsymbol{\mu}_q) - u + \ln |\mathbf{A}^{-1}| - \ln |\mathbf{A}_q^{-1}|).$$

В качестве приближенного значения интеграла

$$\int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{f}) d\mathbf{w}$$

предлагается использовать формулу

$$\int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{f}) d\mathbf{w} \approx \sum_{i=1}^m \log p(y_i|\mathbf{x}_i, \mathbf{w}_i),$$

где  $\mathbf{w}_i$  — реализация случайной величины из распределения  $q(\mathbf{w})$ .

Итоговая функция оптимизации (12) имеет вид

$$(13) \quad \mathbf{f} = \arg \max_{\mathbf{A}_q, \boldsymbol{\mu}_q} \sum_{i=1}^m \log p(y_i|\mathbf{x}_i, \mathbf{w}_i) - D_{\text{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})).$$

*Пример 2.* Пусть задана выборка  $\mathfrak{D}$ , в которой переменная  $y$  не зависит от  $\mathbf{x}$ :

$$(14) \quad y \sim \mathcal{N}(\mathbf{w}, \mathbf{B}^{-1}),$$

$$\mathbf{B}^{-1} = \begin{pmatrix} 2 & 1,8 \\ 1,8 & 2 \end{pmatrix},$$

$$p(\mathbf{w}|\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

График аппроксимации распределения параметров представлен на рис. 1,а. Как видно из графика, с использованием метода (13) получено грубое приближение апостериорного распределения  $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{f})$ , что может существенно занизить оценку правдоподобия модели.

Данный пример показывает, что качество итоговой аппроксимации распределения  $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{f})$  значительно зависит от схожести распределений  $\hat{q}$  и  $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{f})$ . В силу диагональности матрицы  $\mathbf{A}_q$  и полного ранга матрицы  $\mathbf{B}$  итоговое распределение  $\hat{q}$  не может адекватно приблизить данное распределение  $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{f})$ .

### 3.2. Аппроксимация с использованием градиентного метода

В качестве множества распределений  $Q = \{q(\mathbf{w})\}$ , аппроксимирующих неизвестное распределение  $\log p(\mathbf{y}|\mathbf{X}, \mathbf{f})$ , используются распределения параметров, полученные в ходе их оптимизации.

Представим неравенство (8)

$$(15) \quad \log p(\mathbf{y}|\mathbf{X}, \mathbf{f}) \geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{f})}{q(\mathbf{w})} d\mathbf{w} = \mathbb{E}_{q(\mathbf{w})}(\log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{f})) - S(q(\mathbf{w})),$$

где  $S$  — энтропия распределения:

$$S(q(\mathbf{w})) = - \int_{\mathbf{w}} q(\mathbf{w}) \log q(\mathbf{w}) d\mathbf{w},$$

$$p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{f}) = p(\mathbf{w}|\mathbf{f})p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{f}),$$

$\mathbb{E}_{q(\mathbf{w})}(\log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{f}))$  — матожидание логарифма вероятности  $\log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{f})$ :

$$\mathbb{E}_{q(\mathbf{w})}(\log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{f})) = \int_{\mathbf{w}} \log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{f}) q(\mathbf{w}) d\mathbf{w}.$$

Оценка распределений производится при оптимизации параметров. Оптимизация выполняется в режиме мультистарта [11], т.е. при запуске оптимизации параметров модели из нескольких разных начальных приближений. Основная проблема такого подхода — вычисление энтропии  $S$  распределений  $q(\mathbf{w}) \in Q$ . Ниже представлен метод получения оценок энтропии (19)  $S$  и оценок правдоподобия (15).

Запустим  $r$  процедур оптимизаций модели  $\mathbf{f}$  из разных начальных приближений:

$$L(\mathbf{w}^1, \mathbf{y}, \mathbf{X}), \dots, L(\mathbf{w}^r, \mathbf{y}, \mathbf{X}) \rightarrow \min,$$

где  $r$  — число оптимизаций,  $L$  — оптимизируемая функция потерь

$$(16) \quad L = - \sum_{i=1}^m \log p(y_i, \mathbf{w}|\mathbf{x}_i, \mathbf{f}) = -\log p(\mathbf{w}|\mathbf{f}) - \sum_{i=1}^m \log p(y_i|\mathbf{x}_i, \mathbf{w}, \mathbf{f}).$$

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

$$\mathbf{w}^1, \dots, \mathbf{w}^r \sim q^0(\mathbf{w}).$$

Для описания произвольного градиентного метода оптимизации параметров модели введем понятие оператора оптимизации. Оно используется для вычисления оценки энтропии распределения, полученного под действием этой оптимизации.

*Определение 5.* Назовем оператором оптимизации алгоритм  $T$  выбора вектора параметров  $\mathbf{w}'$  по параметрам предыдущего шага  $\mathbf{w}$ :

$$\mathbf{w}' = T(\mathbf{w}).$$

Рассмотрим оператор градиентного спуска:

$$(17) \quad T(\mathbf{w}) = \mathbf{w} - \gamma \nabla L(\mathbf{w}, \mathbf{y}, \mathbf{X}),$$

где  $\gamma$  — длина шага градиентного спуска.

Пусть значения  $\mathbf{w}^1, \dots, \mathbf{w}^r$  — реализации случайной величины из некоторого распределения  $q(\mathbf{w})$ . Начальная энтропия распределения  $q(\mathbf{w})$  соответствует энтропии распределения  $q^0(\mathbf{w})$ , из которого были порождены начальные приближения оптимизации параметров  $\mathbf{w}^1, \dots, \mathbf{w}^r$ . Под действием оператора  $T$  распределение параметров  $\mathbf{w}_1, \dots, \mathbf{w}_r$  изменяется. Для учета энтропии распределений, полученных в ходе оптимизации, формализуем метод, представленный в [4].

*Теорема.* Пусть  $T$  — оператор градиентного спуска,  $L$  — функция потерь, градиент  $\nabla L$  которой имеет константу Липшица  $C_L$ . Пусть  $\mathbf{w}^1, \dots, \mathbf{w}^r$  — начальные приближения оптимизации модели, где  $r$  — число начальных приближений. Пусть  $\gamma$  — длина шага градиентного спуска, такая что

$$(18) \quad \gamma < \frac{1}{C_L}, \quad \gamma < \left( \max_{g \in \{1, \dots, r\}} \lambda_{\max}(\mathbf{H}(\mathbf{w}^g)) \right)^{-1},$$

где  $\lambda_{\max}$  — наибольшее по модулю собственное значение гессиана  $\mathbf{H}$  функции потерь  $L$ .

При выполнении неравенства (18) разность энтропий распределений  $q'(\mathbf{w}), q(\mathbf{w})$  на смежных шагах почти наверное сходится к следующему выражению:

$$(19) \quad S(q'(\mathbf{w})) - S(q(\mathbf{w})) \approx \frac{1}{r} \sum_{g=1}^r (-\gamma \text{Tr}[\mathbf{H}(\mathbf{w}'^g)] - \gamma \text{Tr}[\mathbf{H}(\mathbf{w}'^g)\mathbf{H}(\mathbf{w}'^g)]) + o_{\gamma^2 \rightarrow 0}(1),$$

где  $\mathbf{H}$  — гессиан функции потерь  $L$ .

Получим итоговую формулу для оценки правдоподобия модели.

*Утверждение 2.* Оценка (15) на шаге оптимизации  $\tau$  представима в виде

$$(20) \quad \log \hat{p}(\mathbf{y}|\mathbf{X}, \mathbf{f}) \approx \frac{1}{r} \sum_{g=1}^r L(\mathbf{w}_\tau^g, \mathbf{X}, \mathbf{y}) + S(q^0(\mathbf{w})) + \frac{1}{r} \sum_{b=1}^\tau \sum_{g=1}^r (-\gamma \text{Tr}[\mathbf{H}(\mathbf{w}_b^g)] - \gamma^2 \text{Tr}[\mathbf{H}(\mathbf{w}_b^g)\mathbf{H}(\mathbf{w}_b^g)])$$

с точностью до слагаемых вида  $o_{\gamma^2 \rightarrow 0}(1)$ , где  $\mathbf{w}_b^g$  —  $g$ -я реализация параметров модели на шаге оптимизации  $b$ ,  $q^0(\mathbf{w})$  — начальное распределение.

В [4] предлагается алгоритм приближенного вычисления для выражения, находящегося под знаком суммы в (20):

$$-\gamma \text{Tr}[\mathbf{H}(\mathbf{w}^g)] - \gamma^2 \text{Tr}[\mathbf{H}(\mathbf{w}^g)\mathbf{H}(\mathbf{w}^g)] \approx \mathbf{r}_0^\top (-2\mathbf{r}_0 + 3\mathbf{r}_1 - \mathbf{r}_2),$$

где вектор  $\mathbf{r}_0$  порождается из нормального распределения:

$$\mathbf{r}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{r}_1 = \mathbf{r}_0 - \gamma \mathbf{r}_0^\top \nabla \nabla L, \quad \mathbf{r}_2 = \mathbf{r}_1 - \gamma \mathbf{r}_1^\top \nabla \nabla L.$$

Заметим, что при приближении параметров модели к точке экстремума оценка правдоподобия устремляется в минус бесконечность в силу постоянно убывающей энтропии. Таким образом, чем ближе градиентный метод приближает параметры модели к точке экстремума, тем менее точной становится оценка правдоподобия модели. Один из методов борьбы с данной проблемой будет представлен в разделе 3.3. Доказательство теоремы и утверждения 2 см. в Приложении.

**Модификация алгоритма оптимизации модели.** В качестве оператора  $T$  предлагается использовать псевдослучайный стохастический градиентный спуск, т.е. градиентный спуск, оптимизирующий параметры  $\mathbf{w}^1, \dots, \mathbf{w}^r$  по некоторой случайной подвыборке  $\hat{\mathbf{X}}, \hat{\mathbf{y}}$ , одинаковой для каждой точки старта  $\mathbf{w}^1, \dots, \mathbf{w}^r$ :

$$(21) \quad T(\mathbf{w}) = \mathbf{w} - \frac{m}{\hat{m}} \gamma \nabla L(\mathbf{w}, \hat{\mathbf{y}}, \hat{\mathbf{X}}),$$

где  $\hat{\mathbf{X}}$  — случайная подвыборка выборки  $\mathbf{X}$ , одинаковая для всех точек мультистарта,  $\hat{\mathbf{y}}$  — соответствующие метки классов,

$$|\hat{\mathbf{X}}| = \hat{m}.$$

Как и версия алгоритма с использованием градиентного спуска (21), основной проблемой модифицированного алгоритма оценки интеграла (10) является грубость аппроксимации исходного распределения  $p(\mathbf{w}|\mathbf{f}, \mathfrak{D})$ .

Рассмотрим пример 2 (14). График аппроксимации распределения  $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{f})$  представлен на рис. 1,б. Как видно из графика, градиентный спуск сходится к mode распределения. При небольшом количестве итераций полученное распределение также слабо аппроксимирует апостериорное распределение. При приближении к точке экстремума снижается вариационная оценка правдоподобия модели, что интерпретируется как возможное начало переобучения [4]. Таким образом, снижение оценки (20) можно использовать как критерий остановки оптимизации модели для снижения эффекта переобучения.

На рис. 1 представлена аппроксимация распределения  $p(\mathbf{w}|\mathbf{Y}, \mathbf{X}, \mathbf{f})$  различными методами: *a*) нормальным распределением с диагональной матрицей ковариаций, *b*) с помощью градиентного спуска, *v*) с помощью стохастической динамики Ланжевена. Точками отмечены параметры модели  $\mathbf{f}$ , полученные в ходе нескольких запусков оптимизации и являющиеся реализациями случайной величины с распределением  $q(\mathbf{w})$ . Нормальное распределение слабо аппроксимирует распределение  $p(\mathbf{w}|\mathbf{Y}, \mathbf{X}, \mathbf{f})$  в силу диагональности матрицы ковариаций. Распределение, полученное с помощью градиентного спуска, слабо аппроксимирует распределение  $p(\mathbf{w}|\mathbf{Y}, \mathbf{X}, \mathbf{f})$ , так как сходится к mode.

### 3.3. Аппроксимация с использованием динамики Ланжевена

Для достижения нижней оценки интеграла (10), более близкой к реальному значению логарифма интеграла (3), чем оценка с использованием градиентного спуска, предлагается использовать стохастическую динамику Ланжевена [12]. Стохастическая динамика Ланжевена представляет собой вариант стохастического градиентного спуска с добавлением гауссового шума:

$$(22) \quad T(\mathbf{w}) = \mathbf{w} - \gamma \nabla L - \frac{m}{\hat{m}} \log p(\hat{\mathbf{y}}|\hat{\mathbf{X}}, \mathbf{w}, \mathbf{f}) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \frac{\gamma}{2} \mathbf{I}),$$

где  $\hat{\mathbf{X}}$  — псевдослучайная подвыборка,  $\hat{\mathbf{y}}$  — соответствующие метки,  $\hat{m}$  — размер подвыборки. Длина шага оптимизации  $\gamma$  удовлетворяет условиям, гарантирующим сходимость алгоритма в стандартных ситуациях [12]:

$$\sum_{\tau=1}^{\infty} \gamma_{\tau} = \infty, \quad \sum_{\tau=1}^{\infty} \gamma_{\tau}^2 < \infty.$$

Для оценки энтропии с учетом шума  $\varepsilon$  предлагается использовать следующее неравенство [13, 14]:

$$\hat{S}(q^\tau(\mathbf{w})) \geq \frac{1}{2} u \log \left( \exp \left( \frac{2S(q^\tau(\mathbf{w}))}{u} \right) + \exp \left( \frac{2S(\varepsilon)}{u} \right) \right),$$

где  $\tau$  — текущий шаг оптимизации,  $S(\mathcal{N}(0, \frac{\gamma}{2}))$  — энтропия нормального распределения,  $\hat{S}(q^\tau(\mathbf{w}))$  — энтропия распределения  $q^\tau$  с учетом добавленного шума  $\varepsilon$ .

В отличие от стохастического градиентного спуска стохастическая динамика Ланжевена сходится к апостериорному распределению параметров  $p(\mathbf{w}|\mathcal{D}, \mathbf{f})$  [12, 15]. График аппроксимации апостериорного распределения с использованием динамики Ланжевена представлен на рис. 1,в. При одинаковом количестве итераций динамика Ланжевена продолжает аппроксимировать апостериорное распределение, в то время как градиентный спуск сходится к моде распределения. Как видно из графика, алгоритм, основанный на стохастической динамике Ланжевена, способен давать более точную вариационную оценку правдоподобия (10). В то же время алгоритм более требователен к настройке параметров оптимизации [16]: “быстро изменяющаяся кривизна [траекторий параметров модели] делает методы стохастической градиентной динамики Ланжевена по умолчанию неэффективными”.

#### 4. Вычислительный эксперимент

Для анализа свойств предложенного критерия субоптимальности в задачах регрессии и классификации, а также методов получения нижних оценок правдоподобия модели в задачах выбора моделей был проведен ряд вычислительных экспериментов на выборках Boston Housing, Protein Structure, а также на небольшой подвыборке YearPredictionMSD (далее — Boston, Protein и MSD) [17] и подвыборке изображений рукописных цифр MNIST [18].

Для выборок Boston, Protein и MSD была рассмотрена задача регрессии

$$\mathbf{y} = \mathbf{f}(\mathbf{X}, \mathbf{w}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{f} \in \mathfrak{F}.$$

В качестве множества моделей  $\mathfrak{F}$  были рассмотрены нейросети с одним скрытым слоем и softplus-функцией активации:

$$(23) \quad \mathbf{f}(\mathbf{w}, \mathbf{X}) = \text{softplus}(\mathbf{X}\mathbf{W}_1)\mathbf{W}_2,$$

где  $\mathbf{W}_1 \in \mathbb{R}^{n \times n_1}$  — матрица параметров скрытого слоя нейросети,  $\mathbf{W}_2 \in \mathbb{R}^{n_1 \times 1}$  — матрица параметров выходного слоя нейросети,  $\text{softplus}(\mathbf{X}) = \log(1 + \exp(\mathbf{X}))$ .

Для выборки Boston также было рассмотрено множество моделей с тремя скрытыми слоями, построенных аналогично однослойной модели (23). Размер каждого слоя равнялся 50.

Для выборки MNIST была рассмотрена задача бинарной классификации: из выборки были взяты только объекты, соответствующие цифрам 7 и 9. Размерность выборки была понижена с 784 до 50 методом главных компонент аналогично [19]. Для анализа моделей, полученных в случае высокой вероятности переобучения, из

обучающей выборки были взяты первые 500 объектов. В качестве модели рассматривалась нейросеть с тремя скрытыми слоями

$$f(\mathbf{w}, \mathbf{X}) = \sigma(\text{softplus}(\text{softplus}(\text{softplus}(\mathbf{X}\mathbf{W}_1)\mathbf{W}_2)\mathbf{W}_3)\mathbf{W}_4),$$

где  $\sigma(\mathbf{X}) = (1 + \exp(-\mathbf{X}))^{-1}$  — сигмоида,  $\mathbf{W}_1, \dots, \mathbf{W}_4$  — параметры нейросети.

Во всех экспериментах исходная выборка  $\mathfrak{D}$  разбивалась на обучающую и контрольную подвыборки:  $\mathfrak{D} = \mathfrak{D}_{\text{train}} \sqcup \mathfrak{D}_{\text{test}}$ .

Оптимизация параметров производилась на подвыборке  $\mathfrak{D}_{\text{train}}$ . Для контроля переобучения некоторых алгоритмов из обучающей выборки  $\mathfrak{D}_{\text{train}}$  формировалась валидационная выборка  $\mathfrak{D}_{\text{valid}}$ , на которой не проводилась оптимизация параметров модели. Мощность валидационной выборки  $\mathfrak{D}_{\text{valid}}$  составляла 0,1 мощности обучающей выборки  $\mathfrak{D}_{\text{train}}$ , объекты для валидационной выборки выбирались случайным образом независимо для каждого старта алгоритма. Качество полученных моделей проверялось на подвыборке  $\mathfrak{D}_{\text{test}}$ . Критерием качества модели выступали среднеквадратичное отклонение вектора  $\mathbf{y}$  от вектора  $f(\mathbf{w}, \mathbf{X})$  (RMSE) в случае задачи регрессии и доля верно предсказанных меток класса (Accurasy) в задаче классификации, а также соответствующие критерии при возмущении элементов выборки:

$$(24) \quad \text{RMSE}_{\sigma} = \text{RMSE}(f(\mathbf{w}, \mathbf{X} + \boldsymbol{\varepsilon}), \mathbf{y}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma \mathbf{I}).$$

Были рассмотрены шесть алгоритмов.

1. Базовый алгоритм: оптимизация параметров без валидации и ранней остановки. Оптимизация проводилась с использованием стохастического градиентного спуска (21). Для данного алгоритма априорное распределение  $p(\mathbf{w}|f)$  не использовалось.

2. Алгоритм с валидацией. Для контроля переобучения во время оптимизации качество модели оценивалось на валидационной выборке  $\mathfrak{D}_{\text{valid}}$ . Для данного алгоритма априорное распределение также не использовалось.

3. Алгоритм с валидацией и введенным априорным распределением. В качестве априорного распределения рассматривается распределение вида  $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha \mathbf{I})$ , где  $\alpha$  — дисперсия.

4. Нахождение вариационной нижней оценки с использованием стохастического градиентного спуска.

5. Нахождение вариационной нижней оценки с использованием стохастической динамики Ланжеvena.

6. Нахождение вариационной нижней оценки с аппроксимацией нормальным распределением (13).

Параметры модели выбирались из точек мультистарта (алгоритмы 1–5) или порождались из распределения  $\hat{q}$  (алгоритм 6). Количество точек мультистарта:  $r = 10$  для задач регрессии и  $r = 25$  для задачи классификации. Для алгоритмов 2–6 применялась ранняя остановка: каждые  $\tau_{\text{val}}$  итераций производилась оценка внутреннего критерия качества модели. В качестве критерия остановки применялось следующее условие: значение внутреннего критерия качества не улучшалось  $3\tau_{\text{val}}$  итераций. Для разных алгоритмов внутренним критерием качества выступали различные величины:

1. функция потерь  $L$  (16) на валидационной выборке  $\mathfrak{D}_{\text{valid}}$  для алгоритмов 2, 3,
2. вариационная нижняя оценка правдоподобия (8) на обучающей выборке  $\mathfrak{D}_{\text{train}}$  для алгоритмов 4, 5, 6.

Для каждой модели назначались различные значения параметра  $\alpha$  ( $\alpha \in \{10, \dots, 10^9\}$ ) и длины шага оптимизации  $\gamma$ , отбирались наилучшие модели.

Описание эксперимента представлено в табл. 1. Результаты экспериментов представлены в табл. 2. На рис. 2 представлен график зависимости  $\text{RMSE}_\sigma$  от параметра  $\sigma$  для однослойных моделей.

**Таблица 1. Описание выборок для экспериментов**

Выборка $\mathfrak{D}$	Интервал валидации, $\tau_{\text{val}}$	Количество объектов, $m$	Количество признаков, $n$	Размер подвыборки, $\hat{m}$	Размер скрытого слоя, $n_1$
Boston Housing	100	506	13	$\hat{m} = m$	50
Protein	1000	45000	9	$\hat{m} = 200$	100
MSD	1000	5000	91	$\hat{m} = 50$	100
MNIST	100	500	50	$\hat{m} = 100$	50

**Таблица 2. Результаты эксперимента**

Выборка $\mathfrak{D}$	Алгоритмы					
	1	2	3	4	5	6
Результаты, RMSE/Accuracy						
Boston, один слой	$8,1 \pm 2,0$	$5,9 \pm 0,7$	$5,2 \pm 0,6$	<b><math>3,7 \pm 0,2</math></b>	$6,7 \pm 0,7$	$5,0 \pm 0,4$
Boston, 3 слоя	$7,1 \pm 1,3$	$4,3 \pm 0,1$	$4,4 \pm 0,4$	<b><math>3,2 \pm 0,06</math></b>	$4,6 \pm 0,4$	$6,8 \pm 1,6$
Protein	$5,1 \pm 0,0$	$5,1 \pm 0,0$	$5,1 \pm 0,0$	$5,1 \pm 0,0$	$5,1 \pm 0,0$	<b><math>5,0 \pm 0,1</math></b>
MSD	$12,2 \pm 0,0$	<b><math>10,9 \pm 0,1</math></b>	<b><math>10,9 \pm 0,1</math></b>	$12,2 \pm 0,0$	$12,9 \pm 0,0$	$19,6 \pm 3,6$
MNIST	$0,985 \pm 0,002$	$0,984 \pm 0,002$	<b><math>0,986 \pm 0,002</math></b>	$0,914 \pm 0,005$	$0,979 \pm 0,003$	$0,971 \pm 0,001$
Результаты, RMSE/Accuracy <sub>0,5</sub>						
Boston, один слой	$43,9 \pm 9,4$	$18,6 \pm 2,0$	$15,8 \pm 2,3$	<b><math>11,9 \pm 1,1</math></b>	$20,3 \pm 3,1$	$18,2 \pm 3,3$
Boston, 3 слоя	$23,4 \pm 4,9$	$18,7 \pm 2,8$	$18,3 \pm 3,0$	<b><math>9,0 \pm 0,7</math></b>	$14,5 \pm 2,6$	$15,2 \pm 2,7$
Protein	$19,5 \pm 0,3$	$18,5 \pm 0,5$	$18,6 \pm 0,3$	<b><math>16,7 \pm 0,3</math></b>	$19,3 \pm 0,6$	$19,7 \pm 3,7$
MSD	$178,3 \pm 0,8$	<b><math>121,3 \pm 4,5</math></b>	$123,7 \pm 2,5$	$175,8 \pm 1,0$	$203,8 \pm 1,4$	$292,0 \pm 2,0$
MNIST	$0,931 \pm 0,004$	$0,929 \pm 0,006$	<b><math>0,934 \pm 0,007</math></b>	$0,857 \pm 0,007$	$0,919 \pm 0,008$	$0,916 \pm 0,004$
Результаты, RMSE/Accuracy <sub>1,0</sub>						
Boston, один слой	$120,9 \pm 33,4$	$42,5 \pm 6,3$	$32,5 \pm 6,0$	<b><math>25,7 \pm 3,2</math></b>	$42,4 \pm 5,7$	$41,3 \pm 6,3$
Boston, 3 слоя	$46,1 \pm 15,8$	$40,5 \pm 5,3$	$38,6 \pm 8,0$	<b><math>16,5 \pm 2,5</math></b>	$30,4 \pm 7,9$	$26,2 \pm 6,9$
Protein	$37,0 \pm 0,8$	$34,4 \pm 1,1$	$35,0 \pm 1,0$	<b><math>30,6 \pm 0,6</math></b>	$36,6 \pm 1,1$	$35,0 \pm 8,1$
MSD	$319,6 \pm 1,4$	<b><math>217,5 \pm 8,2</math></b>	$221,9 \pm 4,2$	$314,8 \pm 1,8$	$363,7 \pm 1,9$	$521,6 \pm 3,1$
MNIST	<b><math>0,814 \pm 0,010</math></b>	$0,808 \pm 0,010$	$0,812 \pm 0,008$	$0,772 \pm 0,010$	$0,802 \pm 0,009$	$0,800 \pm 0,009$
Сходимость алгоритмов, тыс. итераций						
Boston, один слой	25	25	25	14	10	27
Boston, 3 слоя	25	4	9	10	1	6
Protein	60	40	80	40	75	85
MSD	250	330	335	250	460	120
MNIST	1	6	3	13	3	25

Модели имеют достаточно большое число параметров, поэтому в ходе оптимизации параметров может произойти переобучение. На выборке Boston Housing базовый алгоритм (1) показал наихудший результат в силу переобучения, при этом алгоритм

4 показал лучший результат по сравнению с алгоритмами 2 и 3. В данном случае использование вариационной оценки предпочтительнее алгоритмов, основанных на кросс-валидации. На выборке Protein все алгоритмы показали схожие результаты. На выборке MSD алгоритмы 4,5,6 показали худший результат в сравнении с алгоритмами, использующими валидационную подвыборку. Наихудший результат показал алгоритм 6, что говорит о значительном отличии апостериорного распределения параметров (4) от нормального.

Алгоритм 6 показал низкое качество (24) при возмущении объектов выборки в большинстве экспериментов. В трех экспериментах наилучшие показатели по данному критерию показал алгоритм 4. Заметим, что алгоритм 5, являющийся модификацией алгоритма 4, показал худшие результаты как по RMSE, так и по RMSE при возмущении объектов выборки. На выборке MNIST алгоритм 4 показал результаты значительно хуже остальных алгоритмов. В целом результаты по данному алгоритму схожи с результатами, описанными в [4]: в отличие от алгоритма 5 алгоритм 4, основанный на стохастическом градиентном спуске, дает заниженную оценку правдоподобия при приближении параметров к точке экстремума. Алгоритм 5, основанный на динамике Ланжевена, также показал худшее время сходимости на выборках MSD и Protein. Возможным дальнейшим улучшением качества этого алгоритма является введение дополнительной корректирующей матрицы, обеспечивающей лучшее время сходления параметров к апостериорному распределению параметров [12].

Программное обеспечение для проведения экспериментов и проверки результатов находится в [20].

## 5. Заключение

В работе были предложены критерии оптимальной и субоптимальной сложности моделей глубокого обучения. Предложен алгоритм выбора субоптимальной модели, основанный на получении вариационной нижней оценки правдоподобия модели. Был предложен метод получения оценки, основанный на стохастическом градиентном спуске, позволяющий проводить выбор модели и оптимизацию модели единообразно. Исследованы свойства стохастического градиентного спуска, а также оценок правдоподобия, полученных с его использованием. Работа представленного алгоритма проиллюстрирована рядом выборок. Вычислительный эксперимент продемонстрировал значимое влияние априорного распределения на апостериорное распределение параметров модели. В силу многоэкстремальности оптимизируемых функций получение аналитических оценок для гиперпараметров модели является вычислительно сложным. В дальнейшем планируется исследовать применение предложенных алгоритмов для оптимизации гиперпараметров градиентными методами, представленными в [8].

## ПРИЛОЖЕНИЕ

*Доказательство утверждения 1.* Доказательство непосредственно следует из (8). Вычитая из обеих частей равенства  $D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{f}))$ , получим

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{f}) - D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{f})) = \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{f})}{q(\mathbf{w})} d\mathbf{w},$$

где  $\log p(\mathbf{y}|\mathbf{X}, \mathbf{f})$  — выражение, не зависящее от  $q(\mathbf{w})$ .

*Доказательство теоремы.* Предварительно приведем две леммы [21, 13], требуемые для доказательства теоремы.

*Лемма 1.* Пусть  $T$  — оператор градиентного спуска,  $L$  — дважды дифференцируемая функция потерь, градиент  $\nabla L$  которой имеет константу Липшица  $C_L$ . Пусть для длины шага  $\gamma$  выполнено неравенство  $\gamma < \frac{1}{C_L}$ . Тогда  $T$  является диффеоморфизмом.

*Лемма 2.* Пусть  $\mathbf{w}$  — случайный вектор с непрерывным распределением  $q(\mathbf{w})$ . Пусть  $T$  — биективное отображение вектора  $\mathbf{w}$  в пространство той же размерности. Пусть  $q'(\mathbf{w})$  — распределение вектора  $T(\mathbf{w})$ . Тогда справедливо утверждение

$$(П.1) \quad S(q'(\mathbf{w})) - S(q(\mathbf{w})) = \int_{\mathbf{w}} q'(\mathbf{w}) \log \left| \frac{\partial T(\mathbf{w})}{\partial \mathbf{w}} \right| d\mathbf{w}.$$

Рассмотрим очередной шаг оптимизации. При  $\gamma < \frac{1}{C}$  оператор градиентного спуска  $T$  является диффеоморфизмом, а значит, и биекцией, справедлива формула (П.1). По усиленному закону больших чисел

$$S(q'(\mathbf{w})) - S(q(\mathbf{w})) \approx \frac{1}{r} \sum_{g=1}^r \log \left| \frac{\partial T(\mathbf{w}'^g)}{\partial \mathbf{w}} \right|.$$

Логарифм якобиана  $\log \left| \frac{\partial T(\mathbf{w}'^g)}{\partial \mathbf{w}} \right|$  оператора  $T$  запишем как

$$(П.2) \quad \log \left| \frac{\partial T(\mathbf{w}'^g)}{\partial \mathbf{w}} \right| = \log |\mathbf{I} - \gamma \mathbf{H}| = \sum_{i=1}^u \log (1 - \gamma \lambda_i),$$

где  $\lambda_i$  —  $i$ -е собственное значение гессиана  $\mathbf{H}$ .

При  $(\gamma \lambda_i)^2 \leq (\gamma \lambda_{\max})^2 < 1$  выражение (П.2) раскладывается в ряд Тейлора:

$$\sum_{t=1}^u \log (1 - \gamma \lambda_i) = -\gamma \text{Tr}[\mathbf{H}(\mathbf{w}'^g)] - \gamma^2 \text{Tr}[\mathbf{H}(\mathbf{w}'^g) \mathbf{H}(\mathbf{w}'^g)] + o_{\gamma^2 \rightarrow 0}(1).$$

Просуммировав полученные выражения для каждой точки мультистарта и вынеся  $o_{\gamma^2 \rightarrow 0}(1)$  за скобки, получим выражение (19), что и требовалось доказать.

*Доказательство утверждения 2.* Представим энтропию распределения  $q^\tau(\mathbf{w})$  следующим образом:

$$S(q^\tau(\mathbf{w})) = S(q^0(\mathbf{w})) - S(q^0(\mathbf{w})) + S(q^1(\mathbf{w})) - S(q^1(\mathbf{w})) + \dots - S(q^{\tau-1}(\mathbf{w})) + S(q^\tau(\mathbf{w})).$$

Каждая разность энтропий вида  $S(q^b(\mathbf{w})) - S(q^{b-1}(\mathbf{w}))$  по теореме с точностью до  $o_{\gamma^2 \rightarrow 0}(1)$  представима в виде

$$(П.3) \quad S(q^b(\mathbf{w})) - S(q^{b-1}(\mathbf{w})) \approx \frac{1}{r} \sum_{g=1}^r (-\gamma \text{Tr}[\mathbf{H}(\mathbf{w}_b^g)] - \gamma^2 \text{Tr}[\mathbf{H}(\mathbf{w}_b^g) \mathbf{H}(\mathbf{w}_b^g)]).$$

Формула (20) получается подстановкой в выражение (15) суммы выражений вида (П.3), а также начальной энтропии  $S(q^0(\mathbf{w}))$ .

## СПИСОК ЛИТЕРАТУРЫ

1. *Grünwald P.* A Tutorial Introduction to the Minimum Description Length Principle // Advances Minimum Descript. Length: Theory Appl. MIT Press, 2005.
2. *Bishop C.* Pattern Recognition and Machine Learning (Information Science and Statistics). Secaucus, USA: Springer-Verlag New York, Inc., 2006.
3. *Graves A.* Practical Variational Inference for Neural Networks // Advances Neural Inform. Proc. Syst. 24 / Ed. by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett et al. Curran Associat. Inc. 2011. P. 2348–2356.
4. *Duvenaud D., Maclaurin D., Adams R.* Early Stopping as Nonparametric Variational Inference // Artific. Intelligen. Statist. 2016. P. 1070–1077.
5. *Salakhutdinov R., Hinton G.* Learning a Nonlinear Embedding by Preserving Class Neighbourhood Structure // J. Machine Learning Res. Proc. Track. 2007. V. 2. P. 412–419.
6. *Sutskever I., Vinyals O., Le Q.* Sequence to Sequence Learning with Neural Networks // Advances Neural Inform. Proc. Syst. 27: Annual Conf. Neural Inform. Proc. Syst. 2014, December 8-13, 2014, Montreal, Quebec, Canada. 2014. P. 3104–3112.
7. *MacKay David J. C.* Information Theory, Inference & Learning Algorithms. USA: Cambridge Univer. Press, 2002.
8. *Maclaurin D., Duvenaud D., Adams R.* Gradient-based Hyperparameter Optimization through Reversible Learning // Proc. 32 Int. Conf. Machine Learning (ICML-15). JMLR Workshop Conf. Proc. 2015.
9. *Hernández-Lobato J. M., Adams R. P.* Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks // Proc. 32 Int. Conf. Machine Learning. 2015. P. 1861–1869.
10. *Kuznetsov M. P., Tokmakova A. A., Strijov V. V.* Analytic and stochastic methods of structure parameter estimation // Informatica. 2016. V. 27. No. 3. P. 607-624.
11. *Shang Y., Wah B.* Global optimization for neural network training // Computer. 1996. Mar. V. 29. No. 3. P. 45–54.
12. *Welling M., Teh Y.* Bayesian Learning via Stochastic Gradient Langevin Dynamics // Proc. 28 Int. Conf. Machine Learning (ICML-11) / Ed. by Lise Getoor, Tobias Scheffer. ICML '11. NY, USA: ACM, 2011. June. P. 681–688.
13. *Dembo A., Cover T., Thomas J.* Information theoretic inequalities // Inform. Theory, IEEE Transact. 1991. V. 37. No. 6. P. 1501–1518.
14. *Altieri N., Duvenaud D.* Variational Inference with Gradient Flows. URL: <http://approximateinference.org/accepted/AltieriDuvenaud2015.pdf>. Дата обращения: 15.03.2017.

15. *Sato I., Nakagawa H.* Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process // Proc. 31 Int. Conf. Machine Learning (ICML-14). 2014. P. 982–990.
16. *Li Chunyuan, Chen Changyou, Carlson David, Carin Lawrence.* Preconditioned Stochastic Gradient Langevin Dynamics for deep neural networks // Proc. Thirtieth AAAI Conf. Artific. Intelligence / AAAI Press. 2016. P. 1788–1794.
17. *Lichman M.* UCI Machine Learning Repository. URL: <http://archive.ics.uci.edu/ml>. Дата обращения: 15.03.2017.
18. *LeCun Yann, Cortes Corinna.* MNIST handwritten digit database. 2010. <http://yann.lecun.com/exdb/mnist/>.
19. *Maclaurin Dougal, Adams Ryan P.* Firefly Monte Carlo: exact MCMC with subsets of data // Proc. 24 Int. Conf. Artific. Intelligence / AAAI Press. 2015. P. 4289–4295.
20. Код вычислительного эксперимента. URL: [svn.code.sf.net/p/mlalgorithms/code/Group074/Bakhteev2016Evidence](http://svn.code.sf.net/p/mlalgorithms/code/Group074/Bakhteev2016Evidence). Дата обращения: 15.03.2017.
21. *Lee J., Simchowitz M., Jordan M., Recht B.* Gradient descent converges to minimizers // Univer. California, Berkeley. 2016. V. 1050. 16 p.

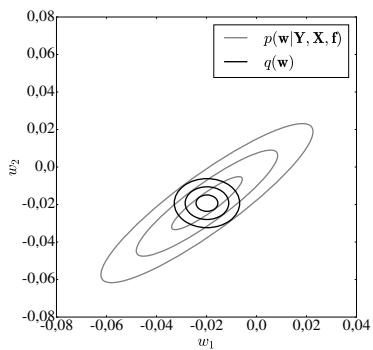
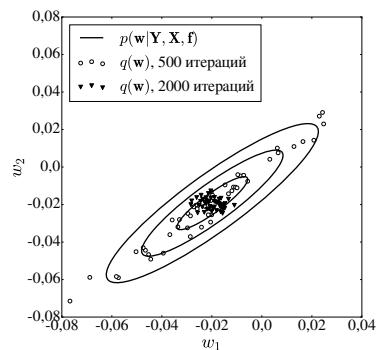
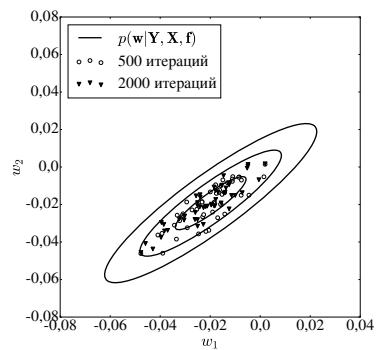
*a**b**c*

Рис. 1

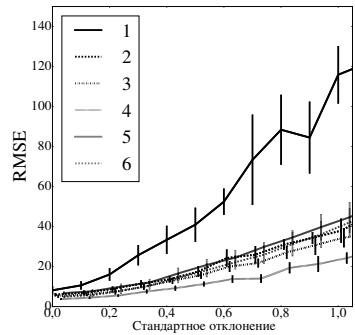
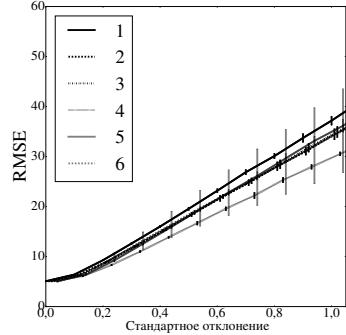
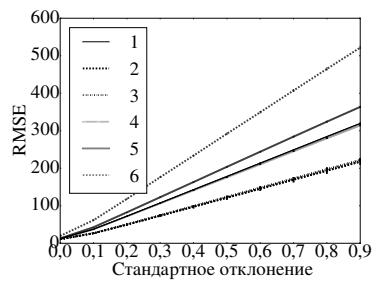
*a**b**c*

Рис. 2

Рис. 1. Аппроксимация распределения *a*) нормальным распределением, *b*) распределением, полученным с помощью градиентного спуска, *c*) с использованием стохастической динамики Ланжевена.

Рис. 2. Возмущение выборки для однослойных нейросетей: *a*) Boston Housing, *b*) Protein, *c*) MSD.