Введение отношения порядка на множестве параметров аппроксимирующих моделей*

А. В. Грабовой[1], **О. Ю. Бахтеев**[2], **В. В. Стрижов**[3]

Аннотация: Исследуется проблема введения отношения порядка на множестве параметров сложных аппроксимирующих моделей. В качестве параметрических моделей исследуются линейные и нейросетевые модели. Порядок на множестве параметров задается при помощи ковариационной матрицы градиентов функции ошибки по параметрам модели. Предлагается использовать заданный порядок для фиксации параметров модели во время решения оптимизационной задачи. Предполагается, что после небольшого числа итераций алгоритма оптимизации некоторые параметры модели можно зафиксировать без значимой потери качества модели. Это позволит существенно понизить размерность задачи оптимизации. В вычислительном эксперименте сравниваются модели, в которых параметры фиксируются в соответствии с предложенным порядком с моделями, в которых параметры фиксируются произвольным образом.

Ключевые слова: аппроксимация выборки; линейная модель; нейросеть; выбор модели; функция ошибки.

DOI: 00.00000/00000000000000

1 Введение

Оптимизация глубоких нейронных сетей является задачей высокой сложности и требует больших вычислительных мощностей [1]. При этом алгоритм оптимизации сходится по большинству параметров сети уже после небольшого числа итераций [2].

*Работа выполнена при поддержке РФФИ (проекты 19-07-1155, 19-07-0875) и НТИ (проект 13/1251/2018).

[1]Московский физико-технический институт, grabovoy.av@phystech.edu

[2]Московский физико-технический институт, bakhteev@phystech.edu

[3]Вычислительный центр имени А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, Московский физико-технический институт, strijov@phystech.edu

Своевременное определения начала сходимости параметров позволит существенно снизить вычислительные затраты на обучения моделей с большим числом параметров.

Примером задания порядка на множестве параметров является l_1–регуляризация [3] и регуляризация ElasticNet [4] для линейных моделей. Порядок, заданный на множестве значений коэффициентов регуляризации, индуцирует порядок на множестве признаковых описаний и указывает на важность признаков. В случае нейросетей для регуляризации параметров используется метод исключения параметров [5, 6]. Данный метод также задает порядок на множестве параметров модели.

В [7] вводится понятие релевантности параметров нейросетевой модели. Оно задает естественный порядок на множестве параметров модели от наименее релевантных до наиболее релевантных. В [8] предложен метод определения релевантности параметров аппроксимирующих моделей при помощи метода Белсли. Релевантность параметров в работе [8] определяется на основе ковариационной матрицы параметров модели.

В данной работе предлагается метод введения отношения порядка на множестве параметров сложных параметрических моделях, таких как нейросеть. Рассматривается порядок, заданный при помощи ковариационной матрицы градиентов функции ошибки по параметрам модели [9]. В работе [2] предложен итерационный метод для поиска ковариационной матрицы градиентов. Данный итерационный метод интегрируется в градиентный метод оптимизации Adam [10].

Множество параметров упорядочивается по возрастанию дисперсии: от параметра с минимальной дисперсией до параметра с максимальной дисперсией градиента функции ошибки по соответствующему параметру модели. Предполагается, что малая дисперсия градиента указывает на то, что соответствующий параметр можно зафиксировать.

Для задания порядка на множестве параметров при помощи ковариационной матрицы вводится предположение о том, что фиксация параметров происходит в момент, когда все параметры модели находятся в некоторой окрестности локального минимума функции ошибки. Данное условие накладывается для корректного использования итерационного метода поиска ковариационной матрицы градиентов.

Заданный порядок на множестве параметров модели используется для фиксации тех параметров модели, которые являются предстоящими с точки зрения заданого порядка. Сначала фиксируются те параметры, которые имеют минимальную дисперсию градиента в окрестности локального минимума функции ошибки.

Для анализа свойств предложенного метода задания порядка на множестве параметров проводится вычислительный эксперимент. В качестве моделей рассматриваются модели различной структурной сложности: линейные модели, нейросетевые модели. Предложенный метод задания порядка сравнивается с методом, в котором порядок задан произвольным образом.

2 Постановка

Задана выборка:

$$\mathfrak{D} = \big\{ \big(\mathbf{x}_i, y_i\big) \big\}_{i=1}^{m}, \quad \mathbf{x}_i \in \mathbb{X} = \mathbb{R}^n, \quad y_i \in \mathbb{Y}, \tag{2.1}$$

где n — размерность признакового пространства, m — количество объектов в выборке. Пространство ответов $\mathbb{Y} = \mathbb{R}$ в случае задачи регрессии и $\mathbb{Y} = \{1, \cdots, K\}$ в случае задачи классификации, где K — количество классов.

Задано семейство моделей, параметрических функций с наперед заданной структурой:

$$\begin{aligned} \mathfrak{F} &= \big\{ f\big(\mathbf{w}\big) : \mathbb{X} \to \mathbb{Y} \mid \mathbf{w} \in \mathbb{R}^p \big\}, \\ \mathbf{h}\big(\mathbf{w}, \mathbf{x}\big) &= \mathbf{W}_1 \boldsymbol{\sigma}\big(\mathbf{W}_2 \boldsymbol{\sigma}\big(\cdots \boldsymbol{\sigma}\big(\mathbf{W}_r \mathbf{x}\big) \cdots \big)\big), \\ f_{\text{cl}}\big(\mathbf{w}, \mathbf{x}\big) &= \arg \max_{j \in \big\{1, \cdots, K\big\}} \text{softmax}\big(\mathbf{h}\big(\mathbf{w}, \mathbf{x}\big)\big)_j, \\ f_{\text{reg}}\big(\mathbf{w}, \mathbf{x}\big) &= \mathbf{h}\big(\mathbf{w}, \mathbf{x}\big), \end{aligned} \tag{2.2}$$

где p — размерность пространства параметров, r — число слоев нейросети, $\mathbf{w} = \text{vec}[\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_r]$, а $\boldsymbol{\sigma}$ — функция активации. В случае задачи регрессии структура модели имеет вид f_{reg}, а в случае классификации имеет вид f_{cl}. Задана функция потерь:

$$\begin{aligned} \mathcal{L}\big(\mathbf{w}, \mathfrak{D}\big) &= \frac{1}{m} \sum_{i=1}^{m} l\big(\mathbf{x}_i, y_i, \mathbf{w}\big), \\ l_{\text{reg}}\big(\mathbf{x}, y, \mathbf{w}\big) &= \big(y - f\big(\mathbf{w}, \mathbf{x}\big)\big)^2, \\ l_{\text{cl}}\big(\mathbf{x}, y, \mathbf{w}\big) &= -\sum_{j=1}^{K} \big([y = j] \ln \text{softmax}_j\big(\mathbf{h}\big(\mathbf{w}, \mathbf{x}\big)\big)\big), \end{aligned} \tag{2.3}$$

где l_{reg} — это функция ошибки на одном элементе для задачи регрессии, l_{cl} — для задачи классификации. Оптимальный вектор параметров $\hat{\mathbf{w}}$ получим минимизацией функции потерь:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^p} \mathcal{L}\big(\mathbf{w}, \mathfrak{D}\big). \tag{2.4}$$

2.1 Задание отношение порядка на множестве параметров

Для поиска оптимальных параметров модели используется градиентный метод оптимизации:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \Delta\mathbf{w}\big(\mathbf{g}_{S,t}, \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \cdots\big), \quad \mathbf{g}_{S,t} = \frac{\partial \mathcal{L}\big(\mathbf{w}_t, \mathbf{X}_S, \mathbf{Y}_S\big)}{\partial \mathbf{w}}, \tag{2.5}$$

где t — номер итерации, $\mathbf{g}_{S,t}$ — значение градиента на подвыборке размера S, $\Delta\mathbf{w}$ — приращение вектора параметров.

Порядок на множестве параметров модели задается при помощи ковариационной матрицы $\mathbf{C}$ градиентов функции ошибки $\mathcal{L}$ по параметрам модели $\mathbf{w}$. Для вычисления

ковариационной матрицы $\mathbf{C}$ используется итерационная формула [2], которая вычисляется на каждой итерации (2.5) градиентного метода оптимизации параметров:

$$\mathbf{C}_t = \big(1 - \kappa_t\big)\mathbf{C}_{t-1} + \kappa_t\big(\mathbf{g}_{1,t} - \mathbf{g}_{S,t}\big)\big(\mathbf{g}_{1,t} - \mathbf{g}_{S,t}\big)^{\mathsf{T}}, \tag{2.6}$$

где t — номер итерации, $\mathbf{g}_{S,t}$ — значение градиента на подвыборке размера S, $\mathbf{g}_{1,t}$ — значение градиента на первом элементе подвыборки, $\kappa_t = \frac{1}{t}$ — параметр сглаживания, $\mathbf{C}s_0$ — инициализируются из равномерного распределения.

Пусть известно t_0 — число итераций, после которого все параметры находятся в некоторой локальной окрестности минимума, тогда, как показано в работе [2], матрица $\mathbf{C}_{t_0}$ аппроксимирует истинную ковариационную матрицу $\mathbf{C}$. Ковариационная матрица $\mathbf{C}_{t_0}$ используется для упорядочивания параметров модели $\mathbf{w}_{t_0}$.

Пусть $\mathcal{I}$ — упорядоченный вектор индексов $[1, 2, \cdots, p]$. Обозначим $\mathcal{I}_{\mathbf{w}_{t_0}}$ — вектор индексов, порядок которого задан при помощи ковариационной матрицы $\mathbf{C}_{t_0}$.

Например, если ковариационная матрица $\mathbf{C}_{t_0}$ имеет вид

$$\begin{bmatrix} 0.3 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.25 \end{bmatrix},$$

тогда вектор индексов $\mathcal{I}_{\mathbf{w}_{t_0}} = [3, 1, 2]$.

2.2 Фиксация параметров

Для фиксации параметров $\mathbf{w}_{t_0}$ при помощи вектора индексов $\mathcal{I}_{\mathbf{w}_{t_0}}$ используется бинарный вектор $\boldsymbol{\alpha}\big(k\big)$:

$$\alpha_i\big(k\big) = \begin{cases} 1 & \text{если } \mathcal{I}_{\mathbf{w}_{t_0}}[j] \leq k, \\ 0 & \text{иначе,} \end{cases} \tag{2.7}$$

где k — число фиксирующих параметров.

Учитывая (2.7) уравнение (2.5) принимает вид

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \boldsymbol{\alpha}\big(k\big) \cdot \Delta\mathbf{w}\big(\mathbf{g}_{S,t}, \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \cdots\big), \tag{2.8}$$

где t — номер итерации, $\mathbf{g}_{S,t}$ — значение градиента на подвыборке размера S, $\Delta\mathbf{w}$ — приращения вектора параметров. После умножения на бинарный вектор $\boldsymbol{\alpha}$, часть параметров не оптимизируются, что приводит к фиксации параметров.

3 Вычислительный эксперимент

Для анализа результатов, полученных предложенным алгоритмом проводится вычислительный эксперимент. В качестве данных использованы синтетические и реальные данные, которые описаны в табл. 1. Выборки MNIST [12] и Boston Housing [11]

Таблица 1: Описание выборок, используемых в эксперименте

Выборка, $\mathfrak{D}$	Тип	Число признаков, n	Модель	Число параметров, p
Boston Housing	регрессия	13	нейросеть	301
MNIST	классификация	784	нейросеть	7960
Synthetic 3	регрессия	200	линейная	200
Synthetic 2	классификация	200	линейная	200
Synthetic 1	регрессия	200	нейросеть	4041

рассматриваются в качестве реальных данных для которых решается задача классификации и регрессии соответсвенно. Синтетические выборки задаются следующим образом:

$$\begin{aligned}\mathfrak{D}_{\text{reg}} =&\big\{\big(\mathbf{x}_i, y_i\big)|\mathbf{x}_i \sim \mathcal{N}\big(\mathbf{0}, \mathbf{I}_n\big),\ y_i \sim \mathcal{N}\big(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i, \mathbf{I}_n\big),\ \mathbf{w} \sim \mathcal{N}\big(\mathbf{0}, \mathbf{I}_n\big)\big\}\\ \mathfrak{D}_{\text{cl}} =&\big\{\big(\mathbf{x}_i, y_i\big)|\mathbf{x}_i \sim \mathcal{N}\big(\mathbf{0}, \mathbf{I}_n\big),\ y_i \sim \mathcal{B}e\big(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i\big),\ \mathbf{w} \sim \mathcal{N}\big(\mathbf{0}, \mathbf{I}_n\big)\big\}\end{aligned} \tag{3.1}$$

В качестве аппроксимирующих моделей рассматриваются линейные и нейросетевые модели (2.2). В качестве функции ошибки для задачи регрессии рассматривается MSELoss, а для задачи классификации CrossEntropyLoss (2.3).

Предварительно для каждой модели и выборки определяется число t_0 — номер итерации после которой все параметры модели находятся в некоторой окрестности локального минимума. Параметр t_0 устанавливается экспериментальным путем для каждой модели и выборки отдельно из условия, что качество модели меняется не значительно при количестве итераций $t > t_0$.

После t_0 шагов алгоритма оптимизации часть параметров модели фиксируются в соответсвии с формулами (2.7), (2.8). Результат работы является усредненным по 25 независимым запускам оптимизации модели. Значение функции ошибки $\mathcal{L}$ усредняется по разным запускам алгоритма оптимизации. В ходе эксперимента проводится анализ вектора $\boldsymbol{\alpha}$, который также усредняется по разным запускам алгоритма оптимизации. Усредненное значение бинарного вектора $\boldsymbol{\alpha}$ обозначим $\hat{\boldsymbol{\alpha}}$.

Выборка Synthetic 1. Эксперимент проводился на синтетически построенных данных. В качестве модели использовалась двухслойная нейросеть — перцептрон. На рис. 1 показаны графики зависимости функции потерь $\mathcal{L}$ от числа фиксируемых параметров. В случае фиксации параметров предложенным методом, функция потерь $\mathcal{L}$ растет медленней чем в случае фиксации параметров произвольным образом.

На рис. 2 показана зависимость векторов $\hat{\boldsymbol{\alpha}}\big(k\big)$ от числа фиксируемых параметров. Каждый столбец соответсвует одному вектору $\hat{\boldsymbol{\alpha}}\big(k\big)$. На рис. 2a, 2c видно, что $\hat{\boldsymbol{\alpha}}\big(k\big)$ имеет большое количество компонент вектора близких к 1. Так как $\hat{\boldsymbol{\alpha}}\big(k\big)$ является усреднением вектора, с компонентами 0 или 1, то предложенный порядок задает некоторый устойчивый порядок на множестве параметров модели. На рис. 2b, 2d видно,

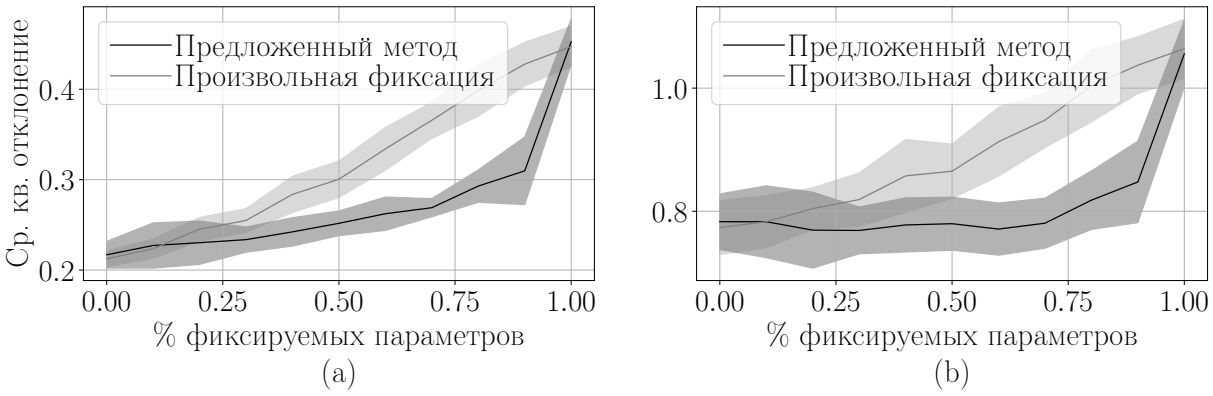

Рис. 1: Зависимость качества модели от количества зафиксированных параметров: (a) на обучающей выборке; (b) на тестовой выборке

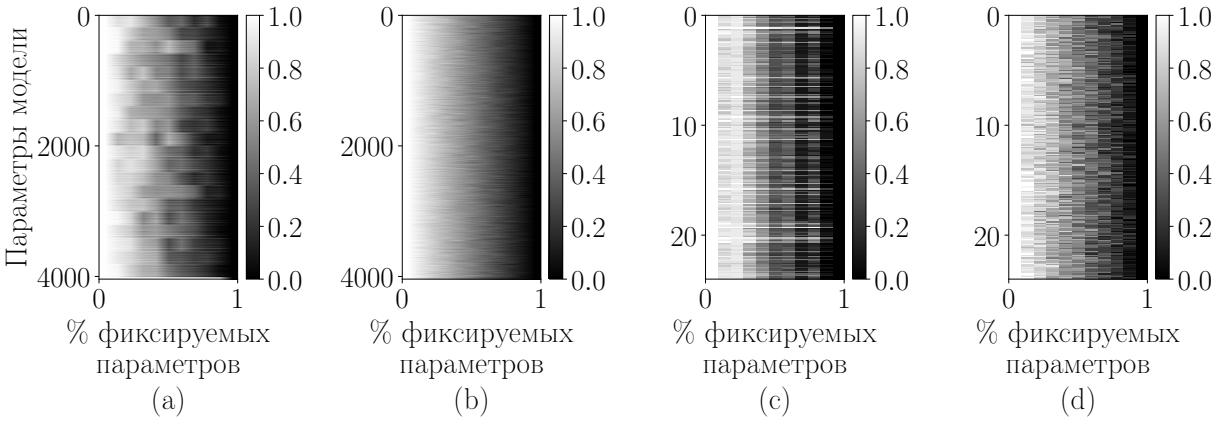

Рис. 2: Визуализация векторов $\hat{\boldsymbol{\alpha}}\big(k\big)$ в зависимости от числа фиксируемых параметров: (a) все параметры модели, упорядочены предложенным методом; (b) все параметры модели, упорядочены произвольным образом; (c) часть параметров модели, упорядочены предложенным методом; (d) часть параметров модели, упорядочены произвольным образом

что в случае произвольной фиксации параметров, компоненты вектора $\hat{\boldsymbol{\alpha}}\big(k\big)$ имеют одинаковые значения, следовательно никакого порядка на множестве параметров нет.

Выборка Boston Housing. Эксперимент проводился на реальных данных. На рис. 3 показаны графики зависимости функции потерь $\mathcal{L}$ от количества фиксируемых параметров. В случае фиксации параметров предложенным методом, функция потерь $\mathcal{L}$ растет также как и в случае фиксации параметров произвольным образом. Данный результат следует из того, что нейросеть является избыточно сложной моделью с большим числом параметров. После фиксации значимого числа параметров у модели остается большое число параметров для дообучения.

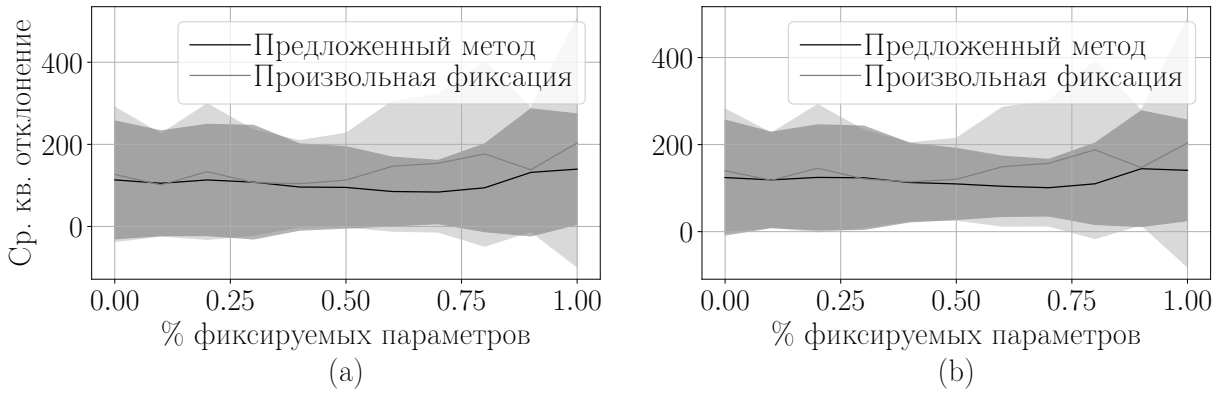

Рис. 3: Зависимость качества модели от количества зафиксированных параметров: a) на обучающей выборке; b) на тестовой выборке

На рис. 4 показана зависимость векторов $\hat{\boldsymbol{\alpha}}(k)$ от числа фиксируемых параметров. На рис. 4a, 4c видно, что $\hat{\boldsymbol{\alpha}}(k)$ меняется не значительно от запуска к запуску алгоритма. Следовательно предложенный порядок задает устойчивый к разным запускам порядок на множестве параметров модели. На рис. 4b, 4d видно, что в случае произвольной фиксации параметров, вектор $\hat{\boldsymbol{\alpha}}(k)$ является произвольным и никакого порядка на множестве параметров нет.

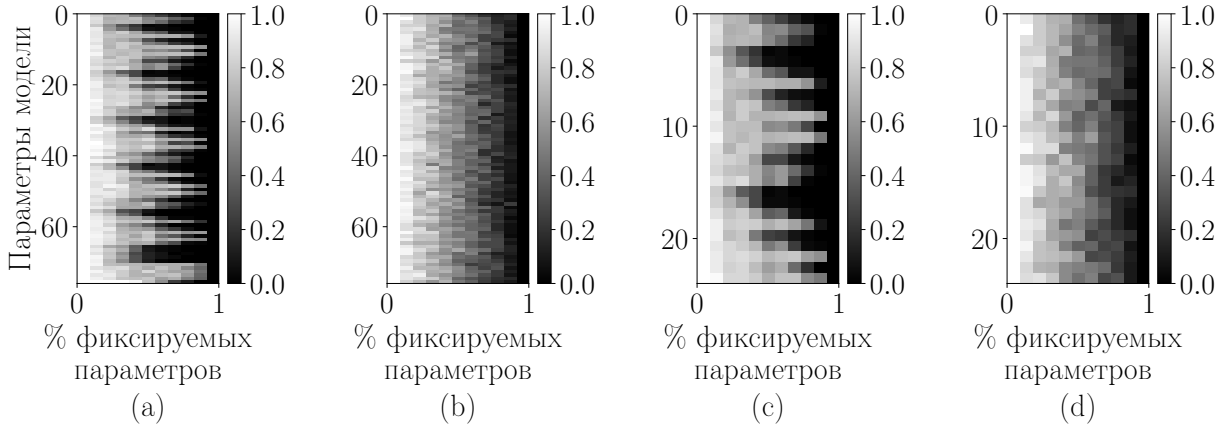

Рис. 4: Визуализация векторов $\hat{\boldsymbol{\alpha}}(k)$ в зависимости от числа фиксируемых параметров: (a) все параметры модели, упорядочены предложенным методом; (b) все параметры модели, упорядочены произвольным образом; (c) часть параметров модели, упорядочены предложенным методом; (d) часть параметров модели, упорядочены произвольным образом

Выборка Synthetic 3. Эксперимент проводился на синтетически построенных данных. В качестве модели использовалась линейная модель регрессии. На рис. 5 показа-

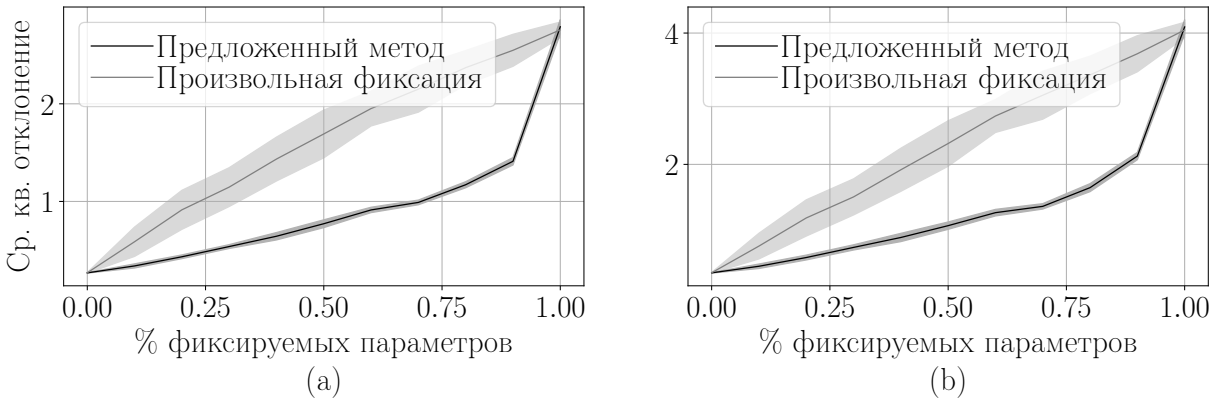

Рис. 5: Зависимость качества модели от количества зафиксированных параметров: (a) на обучающей выборке; (b) на тестовой выборке

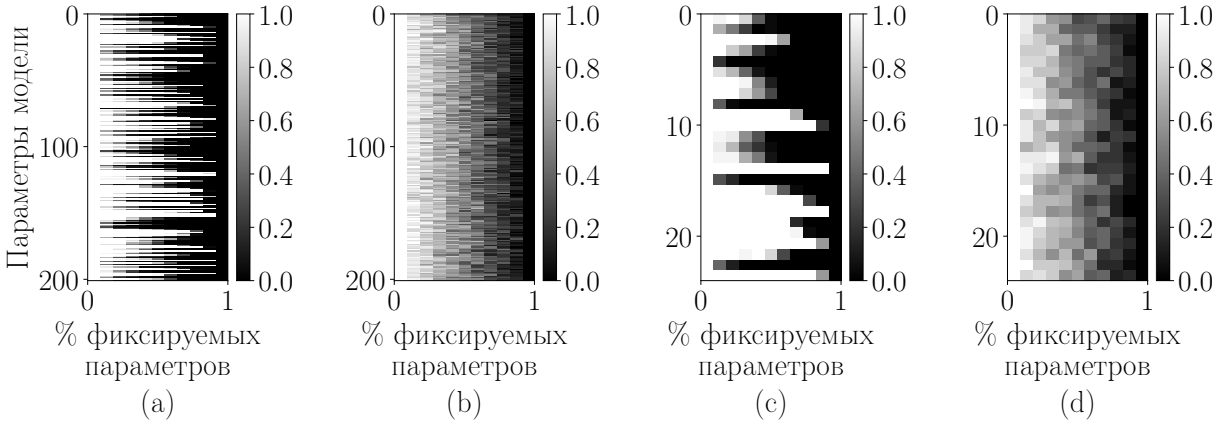

Рис. 6: Визуализация векторов $\hat{\boldsymbol{\alpha}}(k)$ в зависимости от числа фиксируемых параметров: (a) все параметры модели, упорядочены предложенным методом; (b) все параметры модели, упорядочены произвольным образом; (c) часть параметров модели, упорядочены предложенным методом; (d) часть параметров модели, упорядочены произвольным образом

ны графики зависимости функции потерь $\mathcal{L}$ от количества фиксируемых параметров. В случае фиксации параметров предложенным методом функция потерь $\mathcal{L}$ растет значительно медленней в сравнении со случаем фиксации параметров произвольным образом. Дисперсия функции ошибки также значительно меньше в случае фиксации параметров предложенным методом.

На рис. 6 показано, что вектора $\hat{\boldsymbol{\alpha}}(k)$ не меняется от запуска к запуску. Так как данная модель является линейной моделью, следовательно порядок на параметрах модели индуцирует некоторый порядок на множестве признаков.

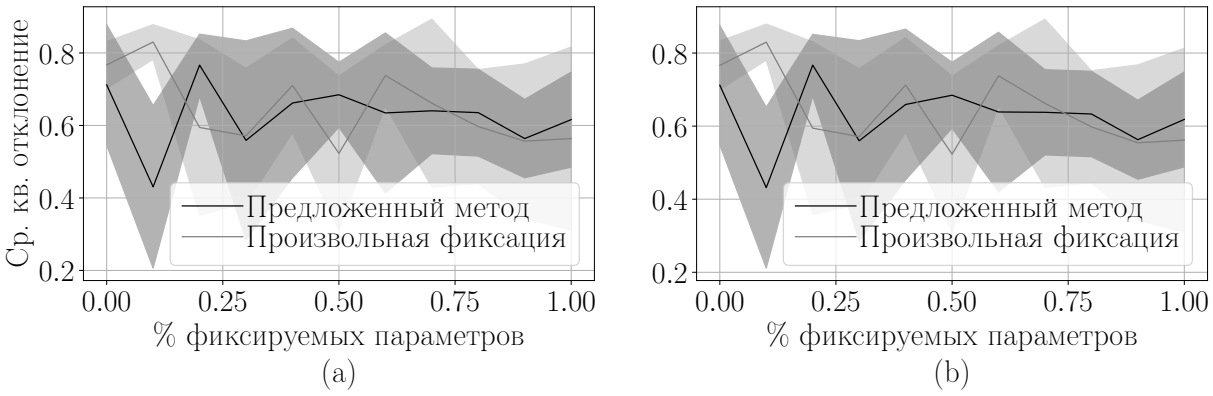

Рис. 7: Зависимость качества модели от количества зафиксированных параметров: (a) на обучающей выборке; (b) на тестовой выборке

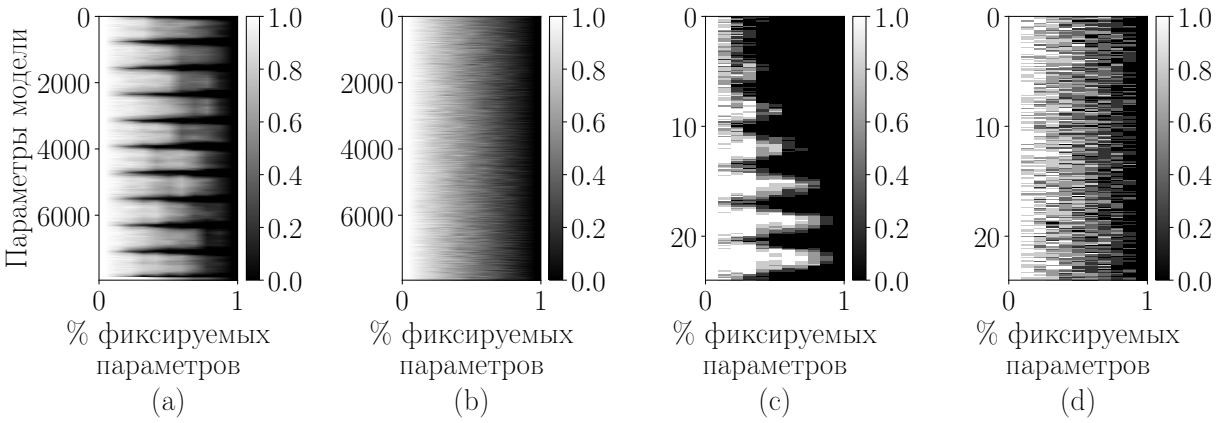

Рис. 8: Визуализация векторов $\hat{\boldsymbol{\alpha}}(k)$ в зависимости от числа фиксируемых параметров: (a) все параметры модели, упорядочены предложенным методом; (b) все параметры модели, упорядочены произвольным образом; (c) часть параметров модели, упорядочены предложенным методом; (d) часть параметров модели, упорядочены произвольным образом

Выборка MNIST. В эксперименте рассматривался двухслойный перцептрон для классификации изображений. В качестве входных данных рассматривались изображения размера 28×28, на которых изображены цифры.

На рис. 7 показано, что график функции ошибки похожи в случае фиксации параметров параметров предложенным методом и в случае произвольной фиксации. Данный результат является следствием того факта, что нейросеть является заведомо переусложненной моделью с большим количеством параметров. После фиксации большого количества параметров у нейросети остается значимое количество параметров модели для дообучения.

На рис. 8 показано, что в случае модели с значимым количеством оптимизацион-

ных параметров, предложенный метод упорядочивания параметров является устойчивым от запуску к запуску.

4 Заключение

В данной работе рассматривалась проблема задания порядка на множестве параметров сложных аппроксимирующих моделей. Рассматривался метод задания порядка на основе анализа стохастических свойств градиента функции ошибки $\mathcal{L}$ по параметрам модели. Для задания порядка использовалась ковариационная матрица градиентов параметров $\mathbf{C}_{\eta_0}$, которая считается итеративно, в течении t_0 итераций градиентного метода параллельно оптимизации. Количество итераций t_0 выбирается заранее экспериментально.

Предложенный метод был проанализирован в вычислительном эксперименте. В эксперименте было показано, что порядок заданный при помощи ковариационной матрицы $\mathbf{C}_{\eta_0}$ является адекватным, так как фиксация параметров в заданном порядке позволяет зафиксировать значимое количество параметров без значимой потери качества.

Отдельно стоит заметить, что параметры упорядычиваюься в процессе оптимизации параметров модели. Как было показано в эксперименте данный порядок является устойчивым и не меняется от запуска к запуску метода оптимизации.

Список литературы

[1] *Sutskever I., Vinyals O., Le Q.* Sequence to Sequence Learning with Neural Networks // Advances in Neural Information Processing Systems, 2014. Vol. 2. P. 3104–3112.

[2] *Li C., Chen C., Carlson D., Carin L.* Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks // Thirtieth AAAI Conference on Artificial Intelligence. — Phoenix, USA, 2016. P. 1788–1794.

[3] *Tibshirani R.* Regression shrinkage and selection via the Lasso // Journal of the Royal Statistical Society, 1996. Vol. 58. P. 267–288.

[4] *Zou H., Hastie T.* Regularization and variable selection via the Elastic Net // Journal of the Royal Statistical Society, 2005. Vol. 67. P. 301–320.

[5] *Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting // Journal of Machine Learning Research, 2014. Vol. 15. P. 1929–1958.

[6] *Molchanov D., Ashukha A., Vetrov D.* Variational Dropout Sparsifies Deep Neural Networks // 34th International Conference on Machine Learning. — Sydney, Australia, 2017. Vol. 70. P. 2498–2507.

[7] *LeCun Y., Denker J., Solla S.* Optimal Brain Damage // Advances in Neural Information Processing Systems, 1989. Vol. 2. P. 598–605.

[8] *Грабовой А. В., Бахтеев О. Ю., Стрижов В. В.* Определение релевантности параметров нейросети // Информатика и ее применение, 2019. Т. 13. Вып. 2. С. 62–70.

[9] *Mandt S., Hoffman M., Blei D.* Stochastic Gradient Descent as Approximate Bayesian Inference // Journal Of Machine Learning Research, 2017. Vol. 18. P. 1–35.

[10] *Kingma D., Ba L.* Adam: A Method for Stochastic Optimization // 3rd International Conference on Learning Representations. — San Diego, USA, 2015.

[11] *Harrison D., Rubinfeld D.* Hedonic prices and the demand for clean air // Journal of Environmental Economics and Management, 1991. Vol. 5. P. 81–102. `https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`.

[12] *LeCun Y., Cortes C., Burges C.* The MNIST dataset of handwritten digits, 1998. `http://yann.lecun.com/exdb/mnist/index.html`