Введение отношения порядка на множестве параметров нейросети

Грабовой Андрей Валериевич

Московский физико-технический институт

ММРО-2019, г. Москва

Задача о введении порядка на множестве параметров

Цель: понижение размерности пространства параметров нейросети в течении процедуры оптимизации.

Задачи

1. Предложить метод введения порядка на множестве параметров.
2. Предложить метод фиксации параметров на основе введенного порядка.

Исследуемая проблема

1. Снижение вычислительно сложности процедуры оптимизации параметров и структуры нейросети.

Метод решения

Для задания порядка предлагается использовать ковариационную матрицу градиентов функции ошибки по параметрам модели.

Список литературы

1. *Tibshirani R.* Regression shrinkage and selection via the Lasso // Journal of the Royal Statistical Society, 1996. Vol. 58. P. 267–288.
2. *Zou H., Hastie T.* Regularization and variable selection via the Elastic Net // Journal of the Royal Statistical Society, 2005. Vol. 67. P. 301–320.
3. *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.
4. *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.
5. *Mandt S., Hoffman M., Blei D.* Stochastic Gradient Descent as Approximate Bayesian Inference // Journal Of Machine Learning Research, 2017. Vol. 18. P. 1–35.
6. *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.

Рассмотрение исследуемой проблемы

Пусть задана некоторая дифференцируемая функция $S(\mathbf{w}, \mathfrak{D})$ по параметрам $\mathbf{w}$. Решается задача:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^p} S(\mathbf{w}, \mathfrak{D}),$$

где p — количество параметров, а $\mathfrak{D}$ обучаемая выборка.
Для решении задачи воспользуемся градиентным методом оптимизации:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \Delta\mathbf{w}\big(\mathbf{g}_{D,t}, \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \cdots\big), \quad \mathbf{g}_{D,t} = \frac{\partial S\big(\mathbf{w}_t, \mathfrak{D}_D\big)}{\partial \mathbf{w}},$$

где t — номер итерации, $\mathbf{g}_{D,t}$ — градиент на подвыборке размера D, а $\Delta\mathbf{w}$ — приращения вектора параметра.
Требуется предложить два бинарных вектора $\boldsymbol{\alpha}(t)$ и $\boldsymbol{\beta}(t)$:

$$\mathbf{w}'_t = \boldsymbol{\beta}(t)\left(\mathbf{w}'_{t-1} + \boldsymbol{\alpha}(t)\Delta\mathbf{w}\big(\mathbf{g}'_{D,t}, \mathbf{w}'_{t-1}, \mathbf{w}'_{t-2}, \cdots\big)\right), \quad \mathbf{g}'_{D,t} = \frac{\partial S\big(\mathbf{w}'_t, \mathfrak{D}_D\big)}{\partial \mathbf{w}},$$

где вектор $\boldsymbol{\alpha}(t)$ отвечает за фиксацию параметров, а вектор $\boldsymbol{\beta}(t)$ — за прореживания параметров.
Вектора $\boldsymbol{\alpha}(t)$ и $\boldsymbol{\beta}(t)$ такие, что:

$$S(\mathbf{w}_T, \mathfrak{D}) \lesssim S\left(\mathbf{w}'_T, \mathfrak{D}\right), \quad \boldsymbol{\alpha}(T) = \mathbf{0}, \quad ||\boldsymbol{\beta}(T)|| \ll p.$$

Постановка задачи машинного обучения

Задана выборка:

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

где n — размерность признакового пространства, m — количество объектов. Пространство ответов $\mathbb{Y} = \mathbb{R}$ в случае задачи регрессии и $\mathbb{Y} = \{1, \cdots, K\}$.
Задано семейство параметрических функций с наперед заданной структурой:

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

Задана функция потерь:

$$\mathcal{L}\big(\mathbf{w}, \mathfrak{D}\big) = \frac{1}{m} \sum_{i=1}^{m} l\big(\mathbf{x}_i, y_i, \mathbf{w}\big),$$

$$l_{\text{reg}}\big(\mathbf{x}, y, \mathbf{w}\big) = \big(y - f\big(\mathbf{w}, \mathbf{x}\big)\big)^2, \quad l_{\text{cl}}\big(\mathbf{x}, y, \mathbf{w}\big) = -\sum_{j=1}^{K} \big([y = j] \ln \text{softmax}_j\big(\mathbf{h}\big(\mathbf{w}, \mathbf{x}\big)\big)\big).$$

Оптимальный вектор параметров $\hat{\mathbf{w}}$:

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

Задание отношения порядка на множестве параметров

Градиентный метод оптимизации:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \Delta\mathbf{w}\big(\mathbf{g}_{D,t}, \mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \cdots\big), \quad \mathbf{g}_{D,t} = \frac{\partial\mathcal{L}\big(\mathbf{w}_t, \mathbf{X}_D, \mathbf{y}_D\big)}{\partial\mathbf{w}},$$

где $\Delta\mathbf{w}$ — приращение вектора параметров.
Ковариационная матрица стохастического градиента:

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

где t — номер итерации, $\mathbf{g}_{D,t}$ — значение градиента на подвыборке размера D, $\mathbf{g}_{1,t}$ — значение градиента на первом элементе подвыборки, $\kappa_t = \frac{1}{t}$ — параметр сглаживания.

Mandt S., Hoffman M., Blei D., 2017

Пусть известно t_0 — число итераций, после которого все параметры находятся в некоторой локальной окрестности минимума, тогда матрица $\mathbf{C}_{t_0}$ аппроксимирует истинную ковариационную матрицу $\mathbf{C}$.

Диагональ ковариационной матрицы $\mathbf{C}_{t_0}$ используется для упорядочивания параметров модели $\mathbf{w}_{t_0}$.
Пусть $\mathcal{J}$ — упорядоченный вектор индексов $[1, 2, \cdots, p]$. Обозначим $\mathcal{J}_{\mathbf{w}_{t_0}}$ — 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 $\mathcal{J}_{\mathbf{w}_{t_0}} = [3, 1, 2]$.
Для фиксации параметров $\mathbf{w}_{t_0}$ при помощи вектора индексов $\mathcal{J}_{\mathbf{w}_{t_0}}$ используется бинарный вектор $\boldsymbol{\alpha}(k)$:

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

где k — число фиксирующих параметров.
Градиентный метод оптимизации:

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

где $\Delta\mathbf{w}$ — приращение вектора параметров.

Для анализа свойств заданого порядка и качества предложенного метода фиксации параметров проведен вычислительный эксперимент. В эксперименте рассматривались реальные и синтетические выборки.

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

Синтетические выборки задаются следующим образом:

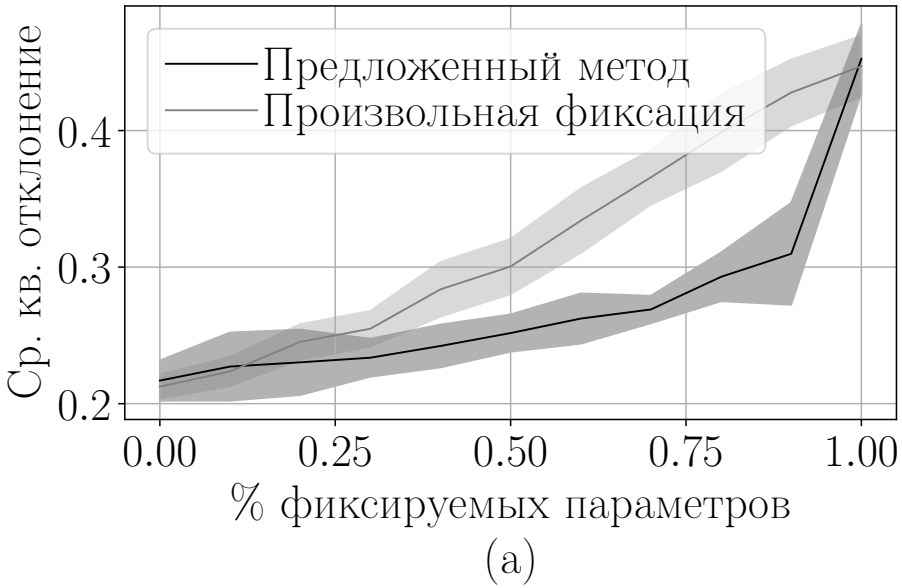
$$\mathfrak{D}_{\text{reg}} = \left\{ \left(\mathbf{x}_i, y_i\right) | \mathbf{x}_i \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_n\right),\ y_i \sim \mathcal{N}\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i, \mathbf{I}_n\right),\ \mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_n\right) \right\}$$

$$\mathfrak{D}_{\text{cl}} = \left\{ \left(\mathbf{x}_i, y_i\right) | \mathbf{x}_i \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_n\right),\ y_i \sim \mathcal{B}e\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i\right),\ \mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_n\right) \right\}$$

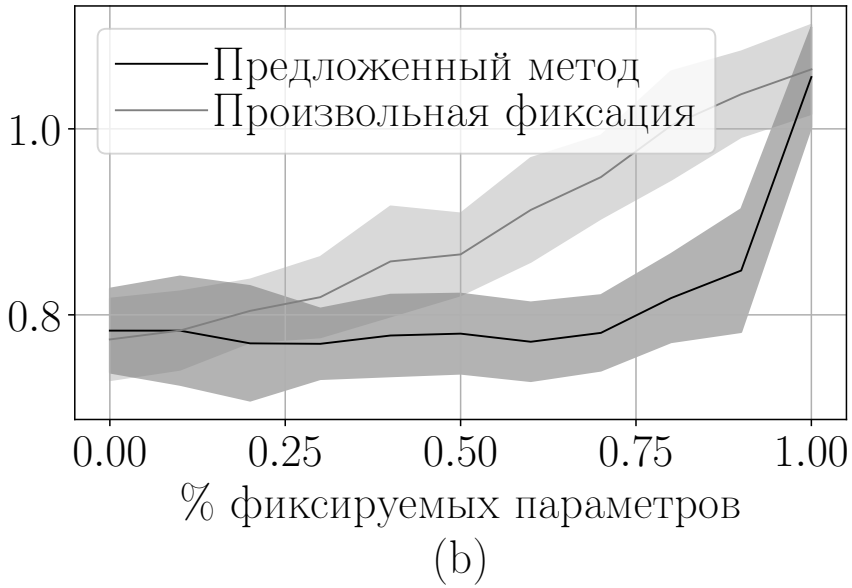
[1] *Harrison D., Rubinfeld D.* Hedonic prices and the demand for clean air // Journal of Environmental Economics and Management, 1991. Vol. 5. P. 81–102.
`https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`.

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

Ошибка модели для выборки Synthetic 1



(a)



(b)

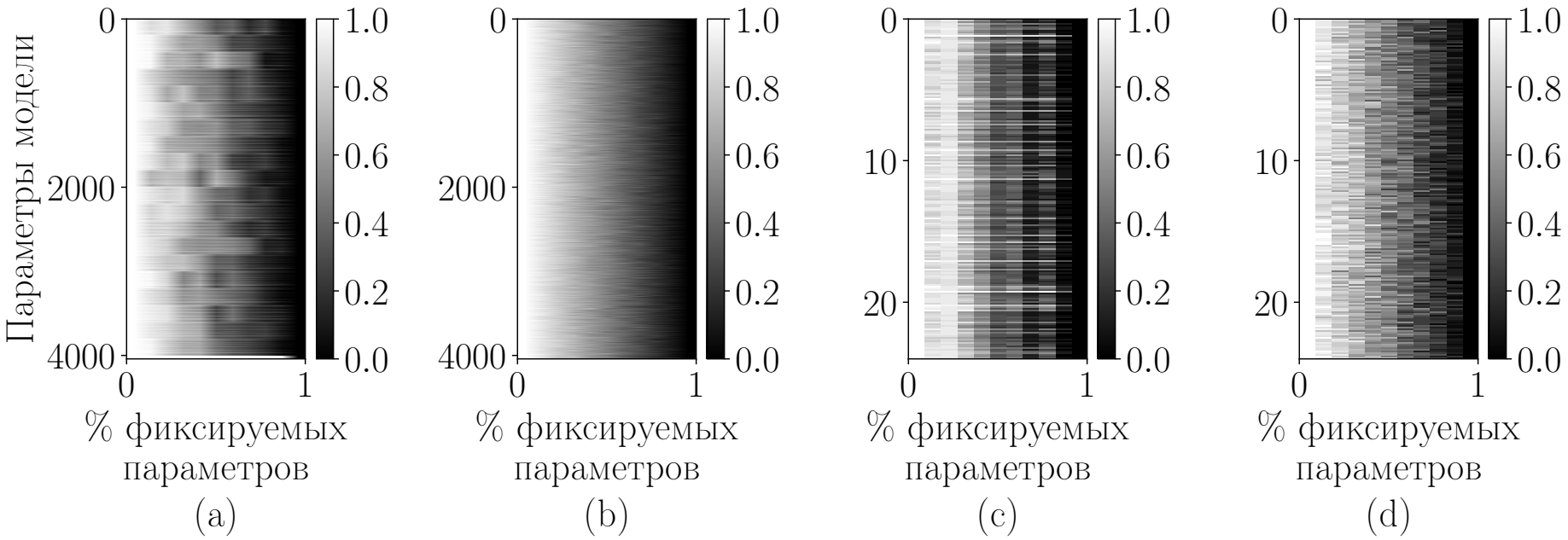
Среднее значение и дисперсия ошибки модели для выборок:

(a) на обучающей выборке;

(b) на тестовой выборке.

Среднее значение функции ошибки растет медленней в случае фиксации параметров в соответсвии с предложенным порядков.

Визуализация векторов $\hat{\boldsymbol{\alpha}}(k)$ для выборки Synthetic 1



Рассматриваются зависимость векторов $\hat{\boldsymbol{\alpha}}(k)$ в зависимости от % фиксируемых параметров в случаях:

(a) все параметры модели; упорядочены предложенным методом;

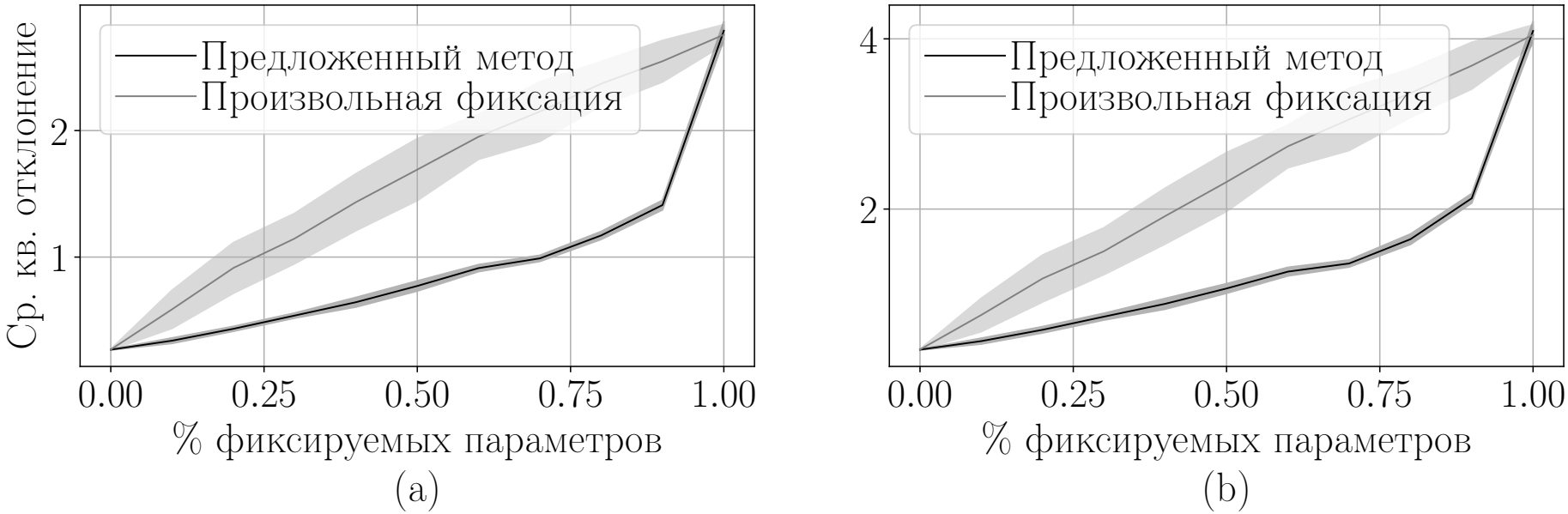
(b) все параметры модели; упорядочены произвольным образом;

(c) часть параметров модели; упорядочены предложенным методом;

(d) часть параметров модели; упорядочены произвольным образом.

Порядок заданный предложенным методом является более устойчивым чем произвольная фиксация параметров модели.

Качество модели для выборки Synthetic 3



(a) на обучающей выборке;

(b) на тестовой выборке.

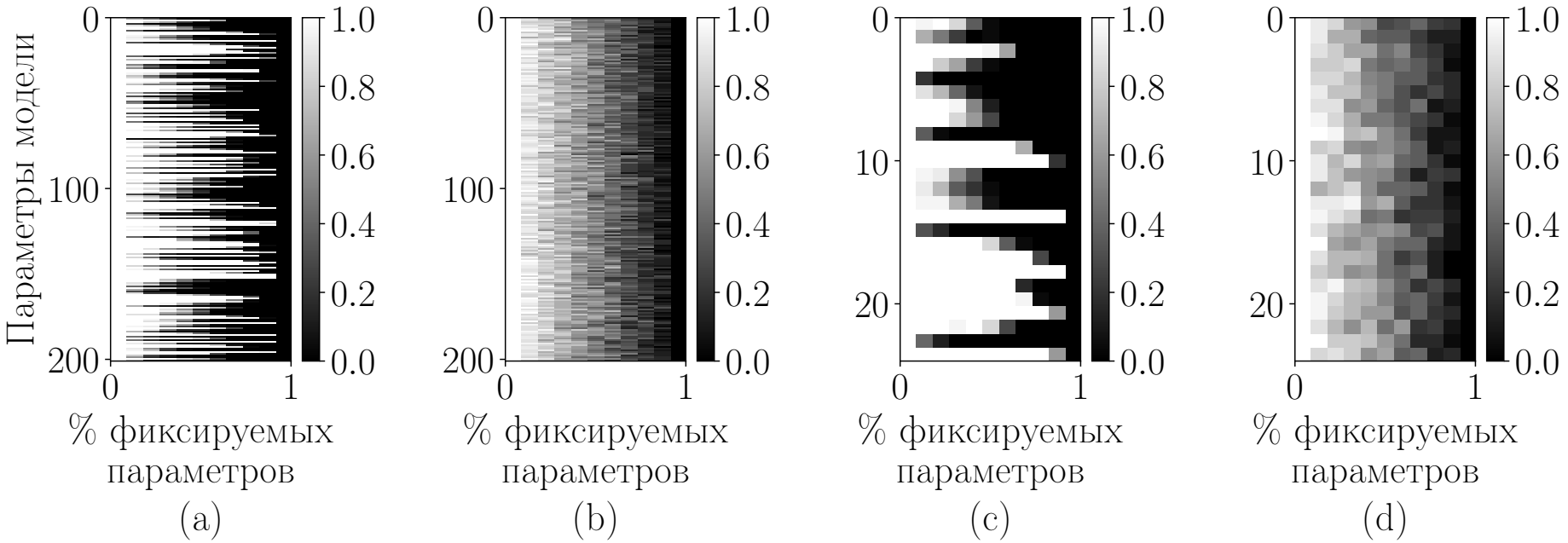
Среднее значение и дисперсия ошибки модели для выборок:

(a) на обучающей выборке;

(b) на тестовой выборке.

Среднее значение функции ошибки растет медленней в случае фиксации параметров в соответсвии с предложенным порядков.

Визуализация векторов $\hat{\boldsymbol{\alpha}}(k)$ для выборки Synthetic 3



Рассматриваются зависимость векторов $\hat{\boldsymbol{\alpha}}(k)$ в зависимости от % фиксируемых параметров в случаях:

(a) все параметры модели; упорядочены предложенным методом;

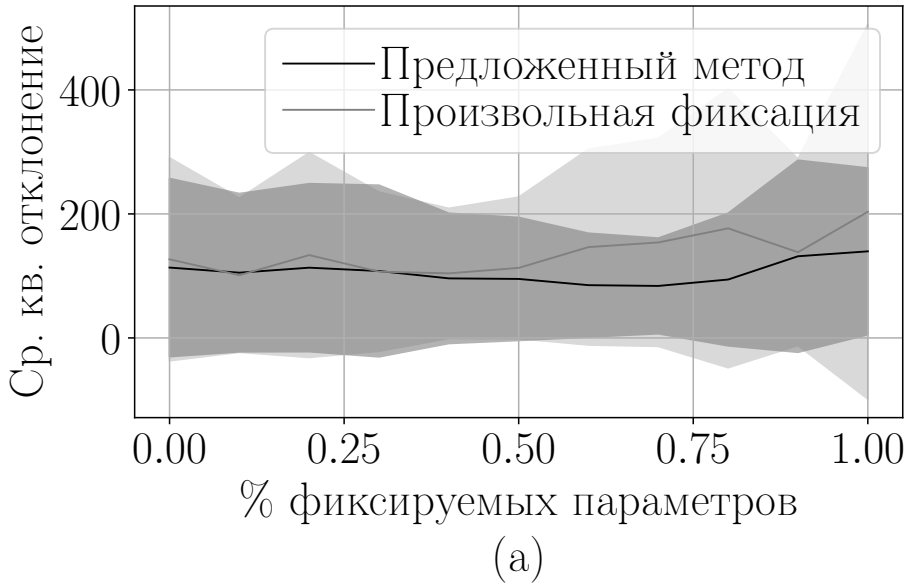
(b) все параметры модели; упорядочены произвольным образом;

(c) часть параметров модели; упорядочены предложенным методом;

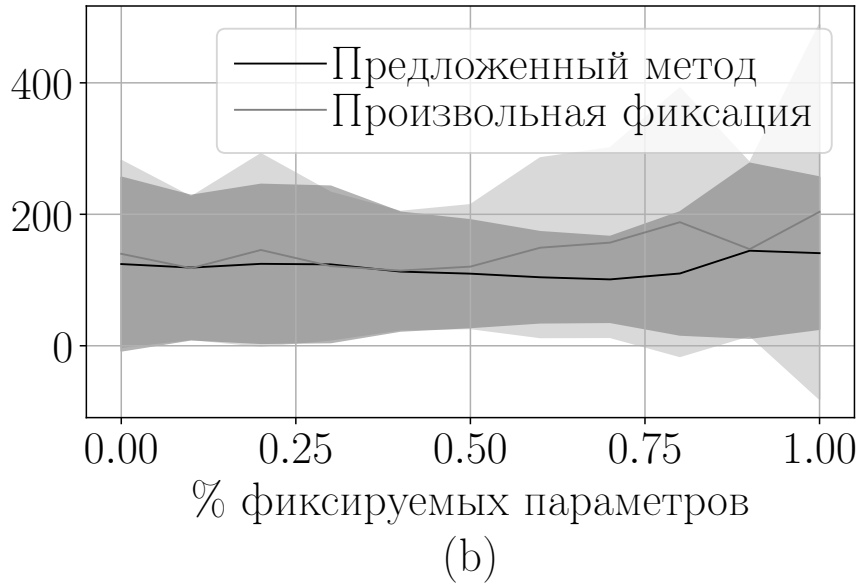
(d) часть параметров модели; упорядочены произвольным образом.

Порядок заданный предложенным методом является более устойчивым чем произвольная фиксация параметров модели.

Качество модели для выборки Boston Housing



(a)



(b)

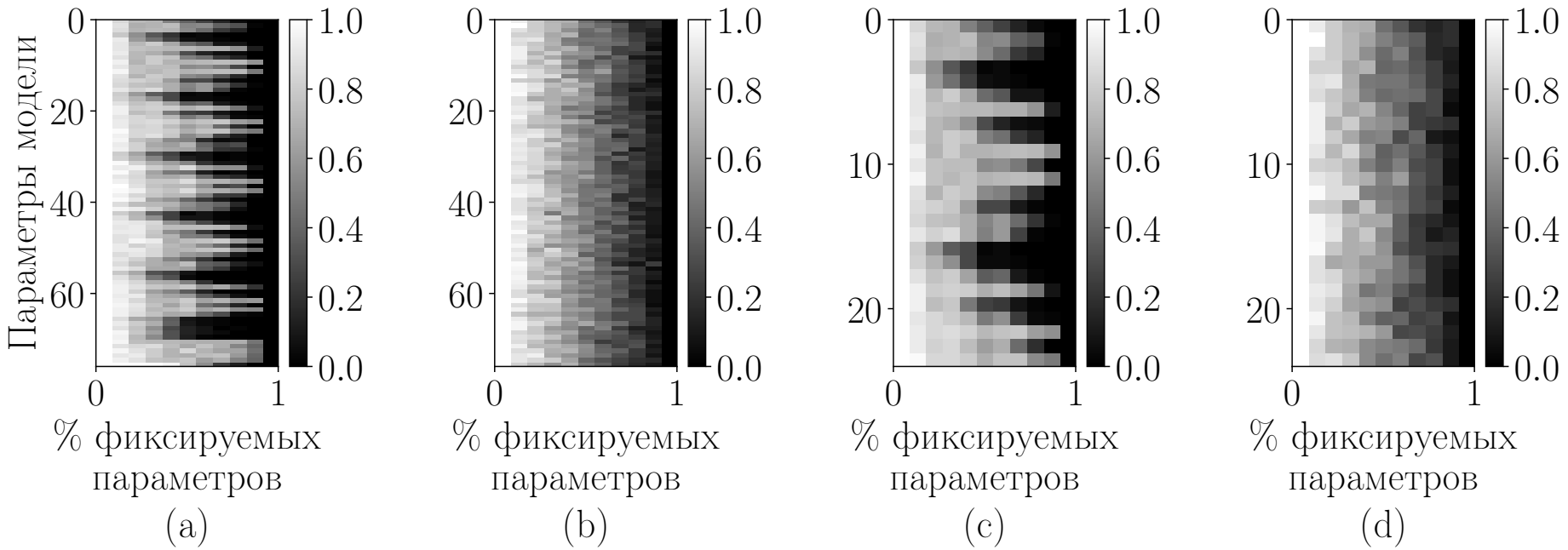
Среднее значение и дисперсия ошибки модели для выборок:

(a) на обучающей выборке;

(b) на тестовой выборке.

Среднее значения функции ошибки имеют сравнимые дисперсии и средние значения в случае фиксации параметров обоими методами.

Визуализация векторов $\hat{\boldsymbol{\alpha}}(k)$ для выборки Boston Housing



Рассматриваются зависимость векторов $\hat{\boldsymbol{\alpha}}(k)$ в зависимости от % фиксируемых параметров в случаях:

(a) все параметры модели; упорядочены предложенным методом;

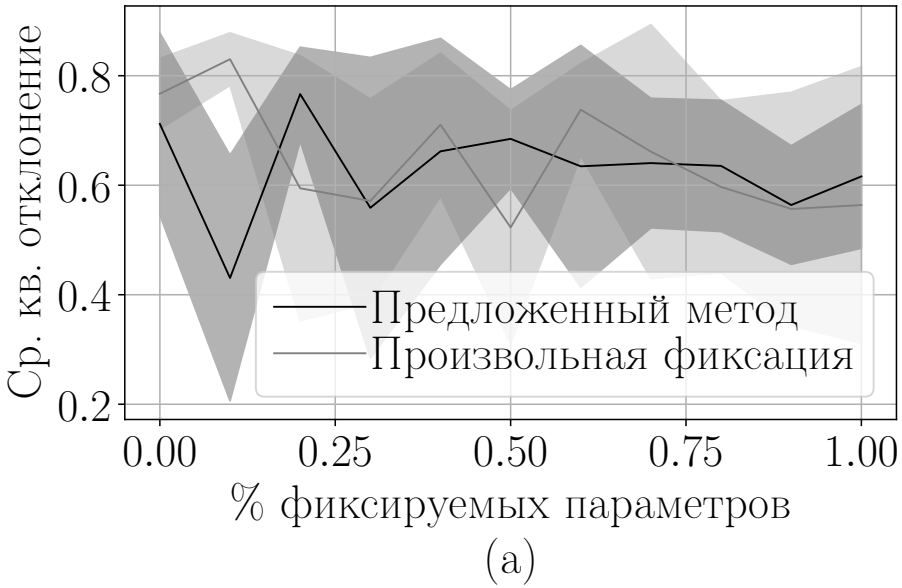
(b) все параметры модели; упорядочены произвольным образом;

(c) часть параметров модели; упорядочены предложенным методом;

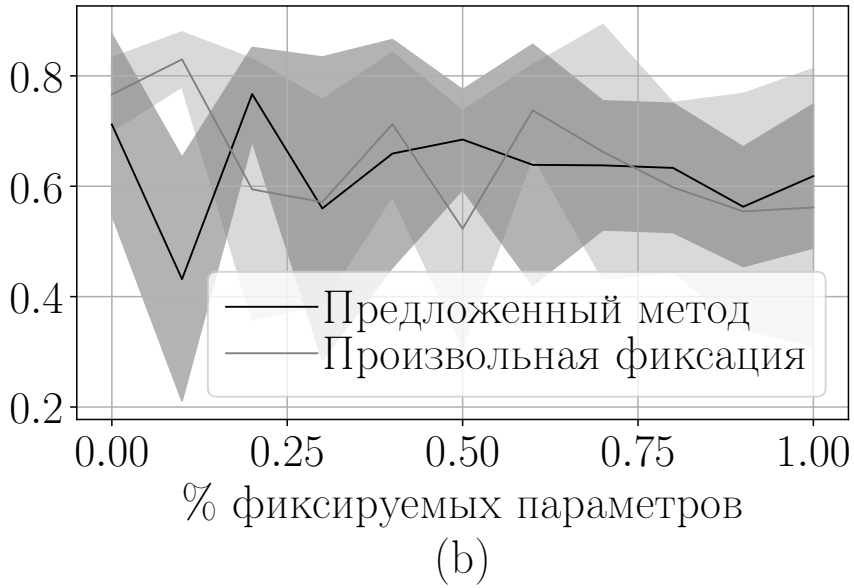
(d) часть параметров модели; упорядочены произвольным образом.

Порядок заданный предложенным методом является более устойчивым чем произвольная фиксация параметров модели.

Качество модели для выборки MNIST



(a)



(b)

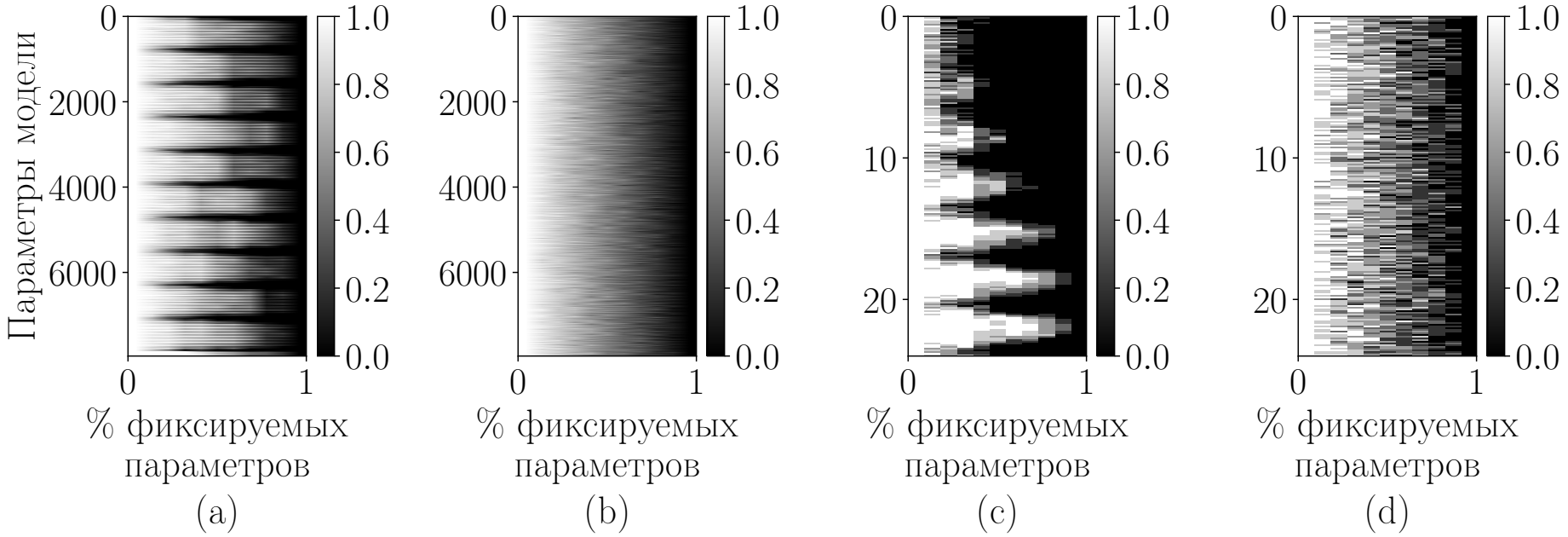
Среднее значение и дисперсия ошибки модели для выборок:

(a) на обучающей выборке;

(b) на тестовой выборке.

Среднее значения функции ошибки имеют сравнимые дисперсии и средние значения в случае фиксации параметров обоими методами.

Визуализация векторов $\hat{\boldsymbol{\alpha}}(k)$ для выборки MNIST



Рассматриваются зависимость векторов $\hat{\boldsymbol{\alpha}}(k)$ в зависимости от % фиксируемых параметров в случаях:

(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.

Порядок заданный предложенным методом является более устойчивым чем произвольная фиксация параметров модели.

Вывод

Сделано:

1. Предложен 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.
2. 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.
3. Также в эксперименте показано, что данный порядок является устойчивым и не меняется от запуска к запуску метода оптимизации.

Планируется:

1. Предлагается исследовать данный метод в применении к transfer learning.
2. Планируется использовать прореживание и фиксацию параметров модели в единой процедуре.