# Дистилляция моделей глубокого обучения

Грабовой Андрей Валериевич

Московский физико-технический институт

МФТИ, г. Долгопрудный

# Вероятностная интерпретация дистилляции моделей

**Цель**
Предложить вероятностную постановку задачи дистилляции моделей глубокого обучения. Развить существующие методы дистилляции и привилегированного обучения использую байесовский подход.

**Задачи**
1. Поставить вероятностную задачу дистилляции для задач классификации и регрессии.
2. Предложить метод байесовской дистилляции нейросетевых моделей.
3. Провести теоретический анализ полученных результатов.

**Исследуемая проблема**
Снижение размерности пространства параметров моделей глубокого обучения.

# Список литературы


1. *Грабовой А.В., Стрижов В.В.* Анализ моделей привилегированного обучения и дистилляции // Автоматика и телемеханика, 2021 (на рецензировании).
2. *Грабовой А.В., Бахтеев О.Ю., Стрижов В.В.* Введение отношения порядка на множестве параметров аппроксимирующих моделей // Информатика и ее применения, 2020, 14(2) : 58-65.
3. *Грабовой А.В., Стрижов В.В.* Байесовская дистилляция моделей глубокого обучения // (текущая работа, в процессе).
4. *Lopez-Paz D., Bottou L., Scholkopf B., Vapnik V.* Unifying Distillation and Privileged Information // In International Conference on Learning Representations. Puerto Rico, 2016.
5. *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // NIPS Deep Learning and Representation Learning Workshop. 2015.
6. *Madala H., Ivakhnenko A.* Inductive Learning Algorithms for Complex Systems Modeling. Boca Raton: CRC Press Inc., 1994.

# Классическая постановка задачи обучения с учителем

Заданы:

1) признаки $\mathbf{x}_i \in \mathbb{R}^n$;
2) привилегированные признаки $\mathbf{x}_i^* \in \mathbb{R}^{n^*}$;
3) целевая переменная $y_i \in \mathbb{Y}$;
4) индексы объектов, для которых известна привилегированная информация $\mathcal{I}$, а для которых она не известна $\bar{\mathcal{I}}$.

Функции учителя $\mathbf{f} : \mathbb{R}^{n^*} \to \mathbb{Y}'$ и ученика $\mathbf{g} : \mathbb{R}^n \to \mathbb{Y}'$ — пространство оценок. Ответ $\mathbf{s}_i = \mathbf{f}\bigl(\mathbf{x}_i^*\bigr)$ функции $\mathbf{f}$ для объекта $\mathbf{x}_i^*$ используется при решении оптимизационной задачи.

Требуется выбрать модель ученика $\mathbf{g}$ из множества

$$\mathfrak{G} = \left\{\mathbf{g} \middle| \mathbf{g} : \mathbb{R}^n \to \mathbb{Y}'\right\}.$$

Оптимизационная задача:

$$\mathbf{g} = \arg\min_{\mathbf{g}\in\mathfrak{G}} \mathcal{L}\bigl(\mathbf{g}, \mathbf{f}, \mathbf{X}, \mathbf{X}^*, \mathbf{y}\bigr),$$

где $\mathcal{L}$ функция ошибки.

# Постановка задачи: дистилляция Хинтона

Заданы:

1) $\mathbf{x}_i^* = \mathbf{x}_i$ для всех $i \in \{1, 2, \cdots, m\}$;
2) $y_i \in \mathbb{Y} = \{1, \cdots, K\}, \quad \mathbb{Y}' = \mathbb{R}^K$.

Параметрические семейства учителя и ученика:

$$\mathfrak{F}_{\text{cl}} = \left\{\mathbf{f} | \mathbf{f} = \text{softmax}\big(\mathbf{v}(\mathbf{x})/T\big), \quad \mathbf{v} : \mathbb{R}^n \to \mathbb{R}^K\right\},$$
$$\mathfrak{G}_{\text{cl}} = \left\{\mathbf{g} | \mathbf{g} = \text{softmax}\big(\mathbf{z}(\mathbf{x})/T\big), \quad \mathbf{z} : \mathbb{R}^n \to \mathbb{R}^K\right\},$$

где $\mathbf{z}, \mathbf{v}$ — это дифференцируемые параметрические функции заданной структуры, $T$ — параметр температуры.

Функция ошибки

$$\mathcal{L}_{\text{st}}(\mathbf{g}) = -\sum_{i=1}^{m} \underbrace{\sum_{k=1}^{K} y_i^k \log \mathbf{g}(\mathbf{x}_i)\Big|_{T=1}}_{\text{исходная функция потерь}} - \sum_{i=1}^{m} \underbrace{\sum_{k=1}^{K} \mathbf{f}(\mathbf{x}_i)\Big|_{T=T_0} \log \mathbf{g}(\mathbf{x}_i)\Big|_{T=T_0}}_{\text{слагаемое дистилляция}},$$

где $\cdot\Big|_{T=t}$ параметр температуры $T$ равняется $t$.

Оптимизационная задача:

$$\hat{\mathbf{g}} = \arg\min_{\mathbf{g} \in \mathfrak{G}_{\text{cl}}} \mathcal{L}_{\text{st}}(\mathbf{g}).$$

# Постановка задачи: дистилляция Вапника

Заданы:

1) $\mathcal{I} = \{1, 2, \cdots, m\}$;
2) $y_i \in \mathbb{Y} = \{1, \cdots, K\}, \quad \mathbb{Y}' = \mathbb{R}^K$.

Параметрическое семейство учителя:

$$\mathfrak{F}^*_{\mathsf{cl}} = \left\{\mathbf{f} \middle| \mathbf{f} = \mathrm{softmax}\bigl(\mathbf{v}^*\bigl(\mathbf{x}^*\bigr)/T\bigr), \quad \mathbf{v}^* : \mathbb{R}^{n^*} \to \mathbb{R}^K\right\},$$

где $\mathbf{v}^*$ — это дифференцируемые параметрические функции заданной структуры, $T$ — параметр температуры.

Функция ошибки:

$$\mathcal{L}_{\mathrm{st}}\bigl(\mathbf{g}\bigr) = -\sum_{i=1}^{m}\sum_{k=1}^{K} y_i^k \log \mathbf{g}\bigl(\mathbf{x}_i\bigr)\bigr|_{T=1} - \sum_{i=1}^{m}\sum_{k=1}^{K} \mathbf{f}\bigl(\mathbf{x}_i^*\bigr)\bigr|_{T=T_0} \log \mathbf{g}\bigl(\mathbf{x}_i\bigr)\bigr|_{T=T_0},$$

где $\cdot\bigr|_{T=t}$ параметр температуры $T$ равняется $t$.

Оптимизационная задача:

$$\hat{\mathbf{g}} = \arg\min_{\mathbf{g} \in \mathfrak{G}_{\mathsf{cl}}} \mathcal{L}_{\mathrm{st}}\bigl(\mathbf{g}\bigr).$$

# Вероятностная постановка

Гипотеза порождения данных:

1) задано распределение целевой переменной $p\big(y_i|\mathbf{x}_i, \mathbf{g}\big)$;
2) задано совместное распределение $p\big(y_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\big)$;
3) для всех $i \in \mathcal{I}$ элементы $y_i$ и $\mathbf{s}_i$ являются зависимыми величинами;
4) если $|\mathcal{I}| = 0$ то решение равно решению максимума правдоподобия.

Совместное правдоподобие истинных меток и меток учителя:

$$p\big(\mathbf{y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}\big) = \prod_{i \not\in \mathcal{I}} p\big(y_i|\mathbf{x}_i, \mathbf{g}\big) \prod_{i \in \mathcal{I}} p\big(y_i, \mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\big).$$

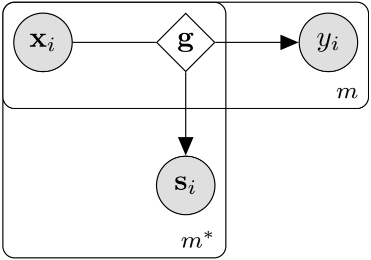
Задача оптимизации:

$$\mathbf{g} = \arg\max_{\mathbf{g} \in \mathfrak{G}} p\big(\mathbf{y}, \mathbf{S}|\mathbf{X}, \mathbf{g}, \mathcal{I}\big),$$

имеет вид:

$$\sum_{i \not\in \mathcal{I}} \log p\big(y_i|\mathbf{x}_i, \mathbf{g}\big) + (1-\lambda) \sum_{i \in \mathcal{I}} \log p\big(y_i|\mathbf{x}_i, \mathbf{g}\big) + \lambda \sum_{i \in \mathcal{I}} \log p\big(\mathbf{s}_i|\mathbf{x}_i, \mathbf{g}\big),$$

где $\lambda \in [0, 1]$ — метапараметр.



# Частный случай: задача классификации

Заданы:

1) учитель $\mathbf{f} \in \mathfrak{F}^*_{\text{cl}}$ и ученик $\mathbf{g} \in \mathfrak{G}_{\text{cl}}$;
2) распределение истинных меток $p\bigl(y|\mathbf{x}, \mathbf{g}\bigr) = \text{Cat}\bigl(\mathbf{g}(\mathbf{x})\bigr)$;
3) распределение ответов учителя $p\bigl(\mathbf{s}|\mathbf{x}, \mathbf{g}\bigr) = C\prod_{k=1}^{K} g_k\bigl(\mathbf{x}\bigr)^{s^k}, \quad C < \infty.$

$$
\hat{\mathbf{g}} = \arg\max_{\mathbf{g}\in\mathfrak{G}} \sum_{i\not\in\mathcal{I}}\sum_{k=1}^{K} y_i^k \log g_k\bigl(\mathbf{x}_i\bigr)\bigr|_{T=1} + (1-\lambda)\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} y_i^k \log g_k\bigl(\mathbf{x}_i\bigr)\bigr|_{T=1}
$$

$$
+ \lambda\sum_{i\in\mathcal{I}}\sum_{k=1}^{K} s_{i,k}\log g_k\bigl(\mathbf{x}_i\bigr)\bigr|_{T=T_0} + \lambda\sum_{i\in\mathcal{I}}\sum_{k=1}^{K}\left(\log g_k\bigl(\mathbf{x}_i\bigr)\bigr|_{T=T_0} + \log\log\frac{1}{g_k\bigl(\mathbf{x}_i\bigr)}\bigr|_{T=T_0}\right)
$$

**Теорема (Грабовой, 2020)**

*Пусть всех $k$ выполняется $1 > 1-\varepsilon > g_k\bigl(\mathbf{x}\bigr) > \varepsilon > 0$, тогда при*

$$
C = (-1)^K \frac{K^{K/2}}{2^{K(K-1)/2}}\prod_{k=1}^{K} g_k\bigl(\mathbf{x}\bigr)\log g_k\bigl(\mathbf{x}\bigr)
$$

*функция $p\bigl(\mathbf{s}|\mathbf{x}, \mathbf{g}\bigr) = C\prod_{k=1}^{K} g_k\bigl(\mathbf{x}\bigr)^{s^k}$ является плотностью распределения.*

# Частный случай: задача регрессии

1) учитель $f \in \mathfrak{F}^*_{\text{rg}} = \{f | f = \mathbf{v}^*(\mathbf{x}^*), \quad \mathbf{v}^* : \mathbb{R}^{n^*} \to \mathbb{R}\}$;
2) ученик $g \in \mathfrak{G}_{\text{rg}} = \{g | g = \mathbf{z}(\mathbf{x}), \quad \mathbf{z} : \mathbb{R}^{n} \to \mathbb{R}\}$;
3) распределение истинных меток $p(y|\mathbf{x}, g) = \mathcal{N}(y|g(x), \sigma)$;
4) распределения меток учителя $p(s|\mathbf{x}, g) = \mathcal{N}(s|g(\mathbf{x}), \sigma_s)$.

Оптимизационная задача:

$$\hat{g} = \arg\min_{g \in \mathfrak{G}} \sum_{i \notin \mathcal{I}} \sigma^2 \left(y_i - g(\mathbf{x}_i)\right)^2 + (1-\lambda) \sum_{i \in \mathcal{I}} \sigma^2 \left(y_i - g(\mathbf{x}_i)\right)^2 + \lambda \sum_{i \in \mathcal{I}} \sigma_s^2 \left(s_i - g(\mathbf{x}_i)\right)^2.$$

## Теорема (Грабовой, 2020)

*Пусть $\mathfrak{G}_{rg}$ — класс линейных функций $g(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x}$. Тогда решение оптимизационной задачи эквивалентно решению задачи линейной регрессии $\mathbf{y}'' = \mathbf{X}\mathbf{w} + \varepsilon,\ \varepsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ , где $\boldsymbol{\Sigma}^{-1} = diag(\boldsymbol{\sigma'})$ и $\mathbf{y}''$ имеют следующий вид:*

$$\sigma'_i = \begin{cases} \sigma^2, \text{ если } i \notin \mathcal{I} \\ (1-\lambda)\sigma^2 + \lambda\sigma_s^2, \text{ иначе,} \end{cases} \qquad \mathbf{y}'' = \boldsymbol{\Sigma}\mathbf{y}', \quad y'_i = \begin{cases} \sigma^2 y_i, \text{ если } i \notin \mathcal{I} \\ (1-\lambda)\sigma^2 y_i + \lambda\sigma_s^2 s_i, \text{ иначе.} \end{cases}$$

# Сводная таблица вычислительного эксперимента

| Dataset | Model | CrossEntropyLoss | Accuracy | StudentSize |
|---|---|---|---|---|
| FashionMnist | without teacher | $0{,}461 \pm 0{,}005$ | $0{,}841 \pm 0{,}002$ | 7850 |
| | with teacher | $0{,}453 \pm 0{,}003$ | $0{,}842 \pm 0{,}002$ | 7850 |
| Synthetic | without teacher | $0{,}225 \pm 0{,}002$ | $0{,}831 \pm 0{,}002$ | 33 |
| | with teacher | $0{,}452 \pm 0{,}001$ | $0{,}828 \pm 0{,}001$ | 33 |
| Twitter | without teacher | $0{,}501 \pm 0{,}006$ | $0{,}747 \pm 0{,}005$ | 1538 |
| | with teacher | $0{,}489 \pm 0{,}003$ | $0{,}764 \pm 0{,}004$ | 1538 |

В таблице показаны результаты вычислительного эксперимента для разных выборок. Точность аппроксимации выборки учеником улучшается при использовании модели учителя при обучении.

# Байесовская дистилляция полносвязной сети

Задана модель учителя с фиксированными параметрами в виде суперпозиции отображений:

$$f = \boldsymbol{\sigma} \circ \mathbf{U}_T \circ \boldsymbol{\sigma} \circ \mathbf{U}_{T-1} \circ \cdots \circ \boldsymbol{\sigma} \circ \mathbf{U}_1,$$

где $\mathbf{U}$ матрицы линейны отображений, $\boldsymbol{\sigma}$ нелинейность. Вектор параметров модели учителя:

$$\mathbf{u} = \mathrm{vec}\bigl([\mathbf{U}_T, \mathbf{U}_{T-1}, \cdots \mathbf{U}_1]\bigr).$$

На основе выборки $\{\mathbf{x}_i, y_i\}_{i=1}^m$ и учителя $f$ требуется выбрать модель ученика:

$$g = \boldsymbol{\sigma} \circ \mathbf{W}_L \circ \cdots \circ \boldsymbol{\sigma} \circ \mathbf{W}_1, \quad \mathbf{W}_l \in \mathbb{R}^{n_s \times n_{s-1}},$$

где $\mathbf{W}$, $\boldsymbol{\sigma}$, $\mathbf{w}$ вводятся аналогично учителю. Задача выбора модели $g$ эквивалента задаче оптимизации вектора параметров $\mathbf{w}$.
Оптимизация параметров $\mathbf{w}$ на основе вариационного вывода:

$$\hat{\mathbf{w}} = \arg\min_{q, \mathbf{w}} \mathrm{D}_{\mathrm{KL}}\bigl(q(\mathbf{w})||p(\mathbf{w}|\mathbf{A})\bigr) - \log p\bigl(\mathbf{y}|\mathbf{X}, \mathbf{w}\bigr),$$

где $q(\mathbf{w})$ вариационное распределение параметров $\mathbf{w}$, а $p(\mathbf{w}|\mathbf{A})$ является априорным распределением вектора параметров модели ученика. Предлагается рассмотреть параметры априорного распределения $p(\mathbf{w}|\mathbf{A})$ как функцию от апостериорного распределения $p(\mathbf{u}|\mathbf{X}, \mathbf{y})$.

# Построение априорного распределения

Проблема: в общем случае пространства параметров учителя $\mathbf{u}$ и ученика $\mathbf{w}$ разное.

Рассмотрим апостериорное распределение параметров модели учителя:

$$p(\mathbf{u}|\mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{u}_0, \boldsymbol{\Sigma}_0).$$

**Пространства параметров совпадают**

- число слоев совпадает $L = T$;
- размеры соответствующих слоев совпадают,

тогда $p(\mathbf{w}|\mathbf{A}) = p(\mathbf{w}|\mathfrak{D})$.

**Пространства параметров не совпадают** В данном случае дистиляция проходит в два этапа:

- выполняется сопоставление моделей учителя и ученика;
- апостериорное распределение учителя назначается априорным распределением ученика.

## Определение

*Сопоставление параметрических моделей — изменение структуры модели (одной или нескольких моделей) в результате которого вектора параметров различных моделей принадлежит одному пространству параметров.*

# Отличие в размере скрытого слоя

- число слоев совпадает $L = T$;
- размеры соответствующих слоев не совпадают: $n_l \leq n_t$, где $n_t$ для всех $l \in \{1, \cdots, L\}, t \in \{1, \cdots, T\}$.

Преобразования $t$-го слоя учителя:

$$\phi\bigl(t, \mathbf{u}\bigr) : \mathbb{R}^{\mathsf{p}_{\mathbf{tr}}} \to \mathbb{R}^{\mathsf{p}_{\mathbf{tr}} - 2n_t}$$

описывает удаление одного нейрона из $t$-го слоя. Новый вектор параметров обозначим $\mathbf{u}' = \phi\bigl(t, \mathbf{u}\bigr)$, а выброшенные параметры обозначим $\mathbf{u}''$

## Теорема (Грабовой, 2021)

*Пусть выполняются следующие условия:*

- $p\bigl(\mathbf{u}|\mathbf{X}, \mathbf{y}\bigr) = \mathcal{N}\bigl(\mathbf{u}_0, \boldsymbol{\Sigma}_0\bigr)$;
- *число слоев совпадает* $L = T$;
- *для всех* $t, l$ *таких, что* $t = l$ *выполняется* $n_s \leq n_l,$.

*Тогда* $p\bigl(\mathbf{u}'|\mathbf{X}, \mathbf{y}\bigr)$ *является нормальным распределением:*

$$p\bigl(\mathbf{u}'|\mathbf{X}, \mathbf{y}\bigr) = \mathcal{N}\bigl(\mathbf{u}_0' + \boldsymbol{\Sigma}_0^{','', } \boldsymbol{\Sigma}_0^{''^{-1}} \left(\mathbf{0} - \mathbf{u}_0^{''}\right), \boldsymbol{\Sigma}_0^{'} - \boldsymbol{\Sigma}_0^{','', } \boldsymbol{\Sigma}_0^{''^{-1}} \boldsymbol{\Sigma}_0^{',''}\bigr).$$

# Отличие в числе слоев нейросети

- соответствующие размеры слоев совпадают, $n_t = n_{t-1}$;
- функция активации удовлетворяет свойству $\boldsymbol{\sigma} \circ \boldsymbol{\sigma} = \boldsymbol{\sigma}$.

Преобразования $t$-го слоя учителя:

$$\psi(t) : \mathbb{R}^{\mathsf{p}_{\mathbf{tr}}} \to \mathbb{R}^{\mathsf{p}_{\mathbf{tr}} - n_t n_{t-1}}$$

описывает удаление $t$-го слоя. Новый вектор параметров обозначим $\mathbf{u}' = \phi(t, \mathbf{u})$, а выброшенные параметры обозначим $\mathbf{u}''$

## Теорема (Грабовой, 2021)

*Пусть выполняются следующие условия:*

- $p(\mathbf{u}|\mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{u}_0, \boldsymbol{\Sigma}_0)$;
- *соответствующие размеры слоев совпадают,* $n_t = n_{t-1}$;
- *функция активации удовлетворяет свойству* $\boldsymbol{\sigma} \circ \boldsymbol{\sigma} = \boldsymbol{\sigma}$.

*Тогда* $p(\mathbf{u}'|\mathbf{X}, \mathbf{y})$ *является нормальным распределением:*

$$p(\mathbf{u}'|\mathbf{X}, \mathbf{y}) = \mathcal{N}\left(\mathbf{u}_0' + \boldsymbol{\Sigma}_0^{','' } \boldsymbol{\Sigma}_0^{''-1} \left(\mathbf{i} - \mathbf{u}_0^{''}\right), \boldsymbol{\Sigma}_0^{'} - \boldsymbol{\Sigma}_0^{','' } \boldsymbol{\Sigma}_0^{''-1} \boldsymbol{\Sigma}_0^{',''}\right),$$

*где* $\mathbf{i} = \mathrm{vec}(\mathbf{I})$

# Порядок на множестве параметров

Задание порядка на множестве параметров:

- случайным образом (используется в рамках вычислительного эксперимента)
- на основе метода оптимального прореживания нейросети:

$$\xi = \arg\min_j h_{jj} \frac{u_j^2}{2},$$

  где $h_{jj}$ коэффициент при квадратичном члене в разложении Тейлора функции ошибки по параметрам модели.
- на основе отношения плотности апостериорного распределения параметра к плотности апостериорного распределения параметра к нулю:

$$\xi = \arg\max_j \frac{p\bigl(0|\mathbf{X}, \mathbf{t}\bigr)}{p\bigl(u_j|\mathbf{X}, \mathbf{t}\bigr)},$$

- выбор на основе анализа мультиколиниарности параметров методов Белсли:

$$\xi = \arg\max_j \frac{\lambda_{\max}}{\lambda_j},$$

  где $\lambda$ являются сингулярными числами ковариационный матрицы параметров.

# Вычислительный экспримент

Синтетическая выборка:

$$\begin{aligned}\mathbf{w} &= \left[w_j : w_j \sim \mathcal{N}\left(0, 1\right)\right]_{n\times 1}, \quad \mathbf{X} = \left[x_{ij} : x_{ij} \sim \mathcal{N}\left(0, 1\right)\right]_{m\times n},\\ \mathbf{y} &= \left[y_i : y_i \sim \mathcal{N}\left(\mathbf{x}_i^{\mathsf{T}}\mathbf{w}, \beta\right)\right]_{m\times 1},\end{aligned}$$

где $\beta = 0{,}1$ — уровень шума в данных. Число признаков $n = 10$, для обучения и тестирования было сгенерировано $m_{\text{train}} = 900$ и $m_{\text{test}} = 124$ объекта.
Модель учителя:

$$f\left(\mathbf{x}\right) = \boldsymbol{\sigma} \circ \mathbf{U}_3 \circ \boldsymbol{\sigma} \circ \mathbf{U}_2 \circ \boldsymbol{\sigma} \circ \mathbf{U}_1 \mathbf{x},$$

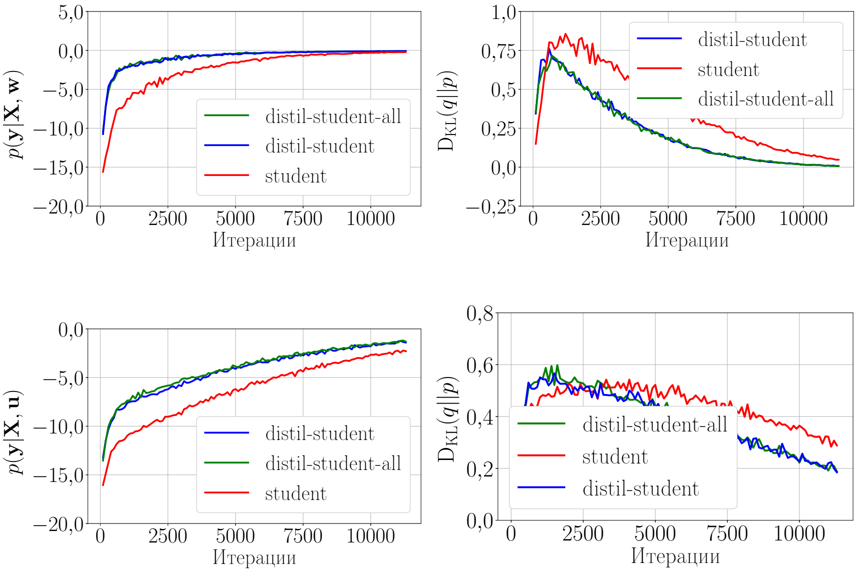
Первая конфигурация модели ученика:

$$g = \boldsymbol{\sigma} \circ \mathbf{W}_3 \circ \boldsymbol{\sigma} \circ \mathbf{W}_2 \circ \boldsymbol{\sigma} \circ \mathbf{W}_1, \quad \mathbf{W}_1 \in \mathbb{R}^{10\times 10}, \mathbf{W}_2 \in \mathbb{R}^{10\times 10}, \mathbf{W}_3 \in \mathbb{R}^{1\times 10}.$$

Вторая конфигурация модели ученика:

$$g = \boldsymbol{\sigma} \circ \mathbf{W}_2 \circ \boldsymbol{\sigma} \circ \mathbf{W}_1, \quad \mathbf{W}_1 \in \mathbb{R}^{1\times 50}, \mathbf{W}_2 \in \mathbb{R}^{50\times 10}.$$

# Результаты экспериментов



# Сводная таблица вычислительного эксперимента

| | teacher | student | distil-student | distil-student-all |
|---|---|---|---|---|
| Эксперимент на синтетической выборке (удаление нейрона) | | | | |
| Архитектура | $[10, 100, 50, 1]$ | $[10, 10, 10, 1]$ | $[10, 10, 10, 1]$ | $[10, 10, 10, 1]$ |
| Число параметров | 6050 | 210 | 210 | 210 |
| Разность площадей | - | 0 | 16559 | 16864 |
| Эксперимент на синтетической выборке (удаление слоя) | | | | |
| Архитектура | $[10, 100, 50, 1]$ | $[10, 50, 1]$ | $[10, 50, 1]$ | $[10, 50, 1]$ |
| Число параметро | | | | |
| Разность площадей | - | 0 | 23310 | 25506 |
| Эксперимент на выборке FashionMnist | | | | |
| Архитектура | $[784, 800, 50, 10]$ | $[784, 50, 10]$ | $[784, 50, 10]$ | $[784, 50, 10]$ |
| Число параметро | | | | |
| Разность площадей | - | 0 | 1165 | 1145 |

# Выносится на защиту

1. Проведен вероятностный анализ задачи дистилляции.
2. Выполнена обобщение классического подхода введя вероятностные предположения о природе данных.
3. Теоретические анализы сформулированы в виде теорем для задачи классификации и регрессии.
4. Поставлена задача байесовской дистилляции моделей глубокого обучения.
5. Предложен метод задания априорного распределения параметров модели ученика на основе апостериорного распределения парамтеров учителя.
6. Доказаны теоремы, которые позволяют проводить сведение структуры модели учителя к структуре модели ученика.
7. Проведен ряд вычислительных экспериментов, которые показывают применимость предложенных методов.

# Публикации ВАК по теме


1. *Грабовой А.В., Бахтеев О.Ю., Стрижов В.В.* Определение релевантности параметров нейросети // Информатика и ее применения, 2019, 13(2).
2. *Грабовой А.В., Бахтеев О. Ю., Стрижов В.В.* Введение отношения порядка на множестве параметров аппроксимирующих моделей // Информатика и ее применения, 2020, 14(2).
3. *A. Grabovoy, V. Strijov.* Quasi-periodic time series clustering for human. Lobachevskii Journal of Mathematics, 2020, 41(3).
4. *Грабовой А.В., Стрижов В.В.* Анализ выбора априорного распределения для смеси экспертов // Журнал Вычислительной математики и математической физики, 2021. 61(5).
5. *Грабовой А.В., Стрижов В.В.* Анализ моделей привилегированного обучения и дистилляции // Автоматика и телемеханика, 2021 (на рецензировании).
6. *T. Gadaev, A. Grabovoy, A. Motrenko, V. Strijov* Numerical methods of minimum sufficient sample size estimation for linear models // in progress.
7. *Базарова А.И., Грабовой А.В., Стрижов В.В.* Анализ свойств вероятностных моделей в задачах обучения с экспертом // подано.
8. *Грабовой А.В., Стрижов В.В.* Байесовская дистилляция моделей глубокого обучения // (текущая работа).