

# Априорное распределение параметров в задачах выбора моделей глубокого обучения

А. В. Грабовой

Диссертация на соискание ученой степени  
кандидата физико-математических наук  
05.13.17 — Теоретические основы информатики  
Научный руководитель: д.ф.-м.н. В. В. Стрижов

21 апреля 2022 г.

# Априорное распределение параметров моделей

Исследуется проблема выбора моделей глубокого обучения. Для выбора моделей используется связный байесовский вывод. Для снижения размерности пространства параметров при выборе модели используется информация об их априорном и апостериорном распределениях.

## Цель исследования —

предложить метод задания априорного распределения параметров модели глубокого обучения с учетом накопленной информации о решаемой задаче.

## Требуется предложить

метод выбора модели, основанный на использовании привилегированной и накопленной информации, выравнивания структур параметрических моделей для снижения размерности пространства параметров моделей.

## Практическая ценность

Снижение размерности пространства параметров моделей глубокого обучения при незначительной потере качества позволяет использовать модели глубокого обучения на устройствах с низкой производительностью, в частности, на мобильных устройствах.

# Привилегированное обучение В. Н. Вапника<sup>1</sup> и дистилляция Дж. Хинтона<sup>2</sup>

Заданы

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

Модели учителя  $\mathbf{f} : \mathbb{R}^{n^*} \rightarrow \mathbb{Y}'$  и ученика  $\mathbf{g} : \mathbb{R}^n \rightarrow \mathbb{Y}'$  — пространство оценок.  
Ответы  $\mathbf{s}_i = \mathbf{f}(\mathbf{x}_i^*)$  функции  $\mathbf{f}$  для объектов  $\mathbf{x}_i^*$ .

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

$$\mathfrak{G} = \{\mathbf{g} | \mathbf{g} : \mathbb{R}^n \rightarrow \mathbb{Y}'\}.$$

Оптимизационная задача:

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

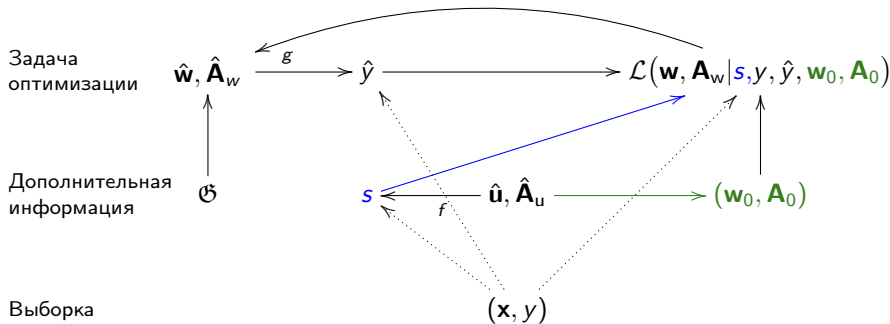
где  $\mathcal{L}$  — функция ошибки.

---

<sup>1</sup>Lopez-Paz D., Bottou L., Scholkopf B., Vapnik V. **Unifying distillation and privileged information** // ICLR, 2016.

<sup>2</sup>Hinton G., Vinyals O., Dean J. **Distilling the knowledge in a neural network** // NIPS, 2015.

# Байесовская дистилляция модели



Выборка  $\mathcal{D} \ni (\mathbf{x}, y), \quad \mathbf{x} \in \mathbb{R}^n, y \in \mathbb{Y}$

Ученик  $\mathcal{G} \ni \mathbf{g} : \mathbb{R}^n \times \mathbb{R}^{n_w} \rightarrow \mathbb{Y}'$  и учитель  $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^{n_w} \rightarrow \mathbb{Y}'$

Прогноз  $\hat{y} = g(\mathbf{x}, \hat{\mathbf{w}})$  и  $\mathbf{s} = f(\mathbf{x}, \hat{\mathbf{u}})$

Априорное распределение  $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0(\hat{\mathbf{u}}, \hat{\mathbf{A}}_u), \mathbf{A}_0(\hat{\mathbf{u}}, \hat{\mathbf{A}}_u)).$

Оптимизационная задача

$$\hat{\mathbf{w}}, \hat{\mathbf{A}}_w = \arg \min_{\mathbf{w} \in \mathbb{R}^{n_w}, \mathbf{A} \in \mathbb{R}^{n_w \times n_w}} \mathcal{L}(\mathbf{w}, \mathbf{A}_w | \mathbf{s}, \mathcal{D}, \mathbf{g}, \mathbf{f}, \mathbf{w}_0, \mathbf{A}_0)$$

# Оптимизация модели ученика на основе учителя

Заданы

- 1) признаки  $\mathbf{x}_i \in \mathbb{R}^n$ ,
- 2)  $y_i \in \mathbb{Y} = \{1, \dots, K\}$ ,  $\mathbb{Y}' = \mathbb{R}^K$ .

Параметрические семейства учителя и ученика:

$$\mathfrak{F}_{\text{cl}} = \{\mathbf{f} | \mathbf{f} = \text{softmax}(\mathbf{v}(\mathbf{x})/T), \quad \mathbf{v} : \mathbb{R}^n \rightarrow \mathbb{R}^K\},$$

$$\mathfrak{G}_{\text{cl}} = \{\mathbf{g} | \mathbf{g} = \text{softmax}(\mathbf{z}(\mathbf{x})/T), \quad \mathbf{z} : \mathbb{R}^n \rightarrow \mathbb{R}^K\},$$

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

Функция ошибки

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

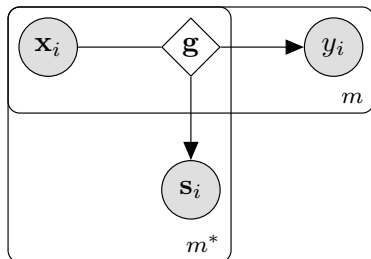
где  $\cdot|_{T=t}$  фиксирует температуру  $T$ .

Оптимальная модель выбирается из класса,  $\hat{\mathbf{g}} = \arg \min_{\mathbf{g} \in \mathfrak{G}_{\text{cl}}} \mathcal{L}(\mathbf{g})$ .

# Вероятностная постановка задачи дистилляции

Гипотеза порождения данных:

- 1) распределение целевой переменной  $p(y_i | \mathbf{x}_i, \mathbf{g})$ ,
- 2) совместное распределение  $p(y_i, \mathbf{s}_i | \mathbf{x}_i, \mathbf{g})$ ,
- 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:

$$p(\mathbf{y}, \mathbf{S} | \mathbf{X}, \mathbf{g}, \mathcal{I}) = \prod_{i \notin \mathcal{I}} p(y_i | \mathbf{x}_i, \mathbf{g}) \prod_{i \in \mathcal{I}} p(y_i, \mathbf{s}_i | \mathbf{x}_i, \mathbf{g}).$$

Задача оптимизации

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

имеет вид

$$\sum_{i \notin \mathcal{I}} \log p(y_i | \mathbf{x}_i, \mathbf{g}) + (1 - \lambda) \sum_{i \in \mathcal{I}} \log p(y_i | \mathbf{x}_i, \mathbf{g}) + \lambda \sum_{i \in \mathcal{I}} \log p(\mathbf{s}_i | \mathbf{x}_i, \mathbf{g}).$$

# Вероятностная постановка задачи классификации

Заданы

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

$$\hat{\mathbf{g}} = \arg \max_{\mathbf{g} \in \mathfrak{G}} \sum_{i \notin \mathcal{I}} \sum_{k=1}^K y_i^k \log g_k(\mathbf{x}_i) \Big|_{T=1} + (1 - \lambda) \sum_{i \in \mathcal{I}} \sum_{k=1}^K y_i^k \log g_k(\mathbf{x}_i) \Big|_{T=1} + \\ + \lambda \sum_{i \in \mathcal{I}} \sum_{k=1}^K s_{i,k} \log g_k(\mathbf{x}_i) \Big|_{T=T_0} + \lambda \sum_{i \in \mathcal{I}} \sum_{k=1}^K \left( \log g_k(\mathbf{x}_i) \Big|_{T=T_0} + \log \log \frac{1}{g_k(\mathbf{x}_i)} \Big|_{T=T_0} \right)$$

## Теорема (Грабовой, 2020)

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

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

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

# Вероятностная постановка задачи регрессии

Заданы

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

Оптимизационная задача:

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

## Теорема (Грабовой, 2020)

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

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



# Байесовская постановка задачи дистилляции

Задана модель учителя, суперпозиция

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

где  $\mathbf{U}$  матрицы линейных отображений,  $\sigma$  монотонная вектор-функция. Параметры учителя фиксированы

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

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

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

где  $\mathbf{W}$ ,  $\sigma$  вводятся как и отображения учителя. Задача выбора модели  $\mathbf{g}$  состоит в оптимизации вектора  $\mathbf{w}$ . Решается вариационным выводом

$$\hat{\mathbf{w}}, \hat{\mu}, \hat{\Sigma} = \arg \min_{\mu, \Sigma, \mathbf{w}} D_{\text{KL}}(q(\mathbf{w}|\mu, \Sigma) || p(\mathbf{w}|\mathbf{A})) - \mathbb{E}_{\mathbf{w} \sim q} \sum_{i=1}^m \log p(y_i | \mathbf{x}_i, \mathbf{w}).$$

Априорное распределение  $p(\mathbf{w}|\mathbf{A})$  задается как функция от апостериорного распределения параметров учителя  $p(\mathbf{u}|\mathbf{X}, \mathbf{y})$ . Оно задано,

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

*Проблема: пространства параметров учителя и ученика не совпадают.*

## Выравнивание структур моделей

Множество структурных параметров задает вид суперпозиции моделей  $\mathbf{f}, \mathbf{g}$ .

### Определение

*Структурные параметры — последовательность размерностей  $n_s$  скрытых представлений после каждого слоя нейросетевой модели.*

### Определение

*Выравнивание структур параметрических моделей — изменение структуры одной или нескольких моделей в результате которого векторы параметров моделей различных структур лежат в одном пространстве.*

### Пространства параметров совпадают

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

тогда

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

### Пространства параметров не совпадают

- выполняется выравнивание моделей учителя  $\mathbf{g}$  и ученика  $\mathbf{f}$ ,
- апостериорное распределение учителя  $p(\mathbf{u}|\mathbf{X}, \mathbf{y})$  назначается априорным распределением ученика  $p(\mathbf{w}|\mathbf{A})$ .

## Размеры скрытых слоев учителя и ученика различны

Преобразования  $t$ -го слоя учителя:

$$\phi(t, \mathbf{u}) : \mathbb{R}^{p_{tr}} \rightarrow \mathbb{R}^{p_{tr} - 2n_t}$$

описывает удаление одного нейрона из  $t$ -го слоя. Новый вектор параметров  $\mathbf{v} = \phi(t, \mathbf{u})$ . Исходный вектор  $\mathbf{u}$  состоит из подвекторов  $\mathbf{v}_2, \mathbf{v}_1, \mathbf{v}$ .

### Теорема (Грабовой, 2021)

Пусть задано распределения вектора параметров  $p(\mathbf{u})$ . Тогда распределение вектора параметров  $p(\mathbf{v})$  представимо в виде:

$$p(\mathbf{v}) = \int_{\mathbf{v}_2 \in \mathbb{R}^{n_t-1}} p(\bar{\mathbf{v}}_1 | \mathcal{D}, \mathbf{v}_1 = \mathbf{0}) d\mathbf{v}_2.$$

### Теорема (Грабовой, 2021)

Пусть выполняются следующие условия:

- 1) апостериорное распределение параметров  $p(\mathbf{u} | \mathcal{D}) = \mathcal{N}(\mathbf{m}, \Sigma)$ ,
- 2) число слоев модели учителя равняется числу слоев модели ученика  $T = L$ ,
- 3) размеры соответствующих слоев не совпадают, другими словами, для всех  $t, l$ , таких что  $t = l$ , выполняется  $n_t \geq n_l$ .

Тогда распределение параметров  $p(\mathbf{v} | \mathcal{D})$  также является нормальным.

## Решение задачи выравнивания структур моделей

$$\hat{\mathbf{w}}, \hat{\mu}, \hat{\Sigma} = \arg \min_{\mu, \Sigma, \mathbf{w}} D_{\text{KL}}(q(\mathbf{w}|\mu, \Sigma) || p(\mathbf{w}|\mathbf{A})) - \mathbb{E}_{\mathbf{w} \sim q} \sum_{i=1}^m \log p(y_i | \mathbf{x}_i, \mathbf{w}).$$

Параметры  $\mathbf{u}$  модели  $\mathbf{f}$  делятся на **удаляемые**  $\nu_2$ , **зануляемые**  $\nu_1$ , оставшиеся  $v$ .  
 Суперпозиция слоев модели учителя  $\mathbf{f}$  в окрестности  $t$ -го слоя:

$$\mathbf{f}(\mathbf{x}) = \cdots \circ \underbrace{\begin{pmatrix} u_{1,1} & \cdots & u_{1,j} & \cdots & u_{1,n_t} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{n_{t+1},1} & \cdots & u_{n_{t+1},j} & \cdots & u_{n_{t+1},n_t} \end{pmatrix}}_{\mathbf{u}_{t+1}} \sigma \circ \underbrace{\begin{pmatrix} u_{1,1} & \cdots & u_{1,n_t-1} \\ \vdots & \ddots & \vdots \\ \color{red}{u_{j,1}} & \cdots & \color{red}{u_{j,n_t-1}} \\ \vdots & \ddots & \vdots \\ u_{n_t,1} & \cdots & u_{n_t,n_t-1} \end{pmatrix}}_{\mathbf{u}_t} \sigma \circ \cdots \circ \mathbf{u}_1 \mathbf{x}$$

Апостериорное распределение параметров  $v$  модели  $\mathbf{f}$ :

$$p(v|\mathcal{D}) = \int_{\color{red}{\nu_2} \in \mathbb{R}^{n_t-1}} p(\bar{\nu}_1|\mathcal{D}, \color{blue}{\nu_1} = \mathbf{0}) d\color{red}{\nu_2}.$$

Из свойства распределения  $p(\bar{\nu}_1|\mathcal{D}, \nu_1 = \mathbf{0}) = \mathcal{N}(\mu, \Xi)$ , с параметрами  $\mu, \Xi$ :

$$\mu = \mathbf{m}_{\bar{\nu}_1} + \Sigma_{\bar{\nu}_1, \nu_1} \Sigma_{\nu_1, \nu_1}^{-1} (\mathbf{0} - \mathbf{m}_{\nu_1}),$$

$$\Xi = \Sigma_{\bar{\nu}_1, \bar{\nu}_1} - \Sigma_{\bar{\nu}_1, \nu_1} \Sigma_{\nu_1, \nu_1}^{-1} \Sigma_{\nu_1, \bar{\nu}_1},$$

Маргинализация нормального распределения  $p(v|\mathcal{D}) = \mathcal{N}(\mu_v, \Xi_{v,v})$ .

## Число скрытых слоев учителя и ученика различны

Преобразования  $t$ -го слоя учителя:

$$\psi(t) : \mathbb{R}^{p_{tr}} \rightarrow \mathbb{R}^{p_{tr} - n_t n_{t-1}}$$

описывает удаление  $t$ -го слоя. Новый вектор параметров  $\mathbf{v} = \psi(t, \mathbf{u})$ , а элементы вектора, которые были удалены как  $\bar{\mathbf{v}}$ .

### Теорема (Грабовой, 2021)

Пусть выполняются следующие условия:

- 1) апостериорное распределение параметров  $p(\mathbf{u}|\mathcal{D}) = \mathcal{N}(\mathbf{m}, \Sigma)$ ,
- 2) соответствующие размеры слоев совпадают,  $n_t = n_{t-1}$ , т.е. матрица  $\mathbf{U}_t$  является квадратной,
- 3) функция активации удовлетворяет свойству идемпотентности  $\sigma \circ \sigma = \sigma$ .

Тогда апостериорное распределение также описывается нормальным распределением с плотностью распределения:

$$p(\mathbf{v}|\mathcal{D}) = \mathcal{N}(\mathbf{m}_v + \Sigma_{v,\bar{v}} \Sigma_{\bar{v},\bar{v}}^{-1} (\mathbf{i} - \bar{\mathbf{v}}), \Sigma_{v,v} - \Sigma_{v,\bar{v}} \Sigma_{\bar{v},\bar{v}}^{-1} \Sigma_{\bar{v},v}),$$

где вектор  $\mathbf{i} = [\underbrace{1, 0, \dots, 0}_{n_t}, \underbrace{0, 1, \dots, 0}_{n_t}, \underbrace{0, 0, 1, \dots, 0}_{n_t}, \underbrace{0, \dots, 1}_{n_t}]^T$ .

## Обобщение для рекуррентной сети RNN

Структура модели RNN задается размерностью скрытых слоев  $n_1, n_2, \dots, n_T$ .  
Отображения  $\phi_{RNN}(t, \mathbf{u}), \psi_{RNN}(t, \mathbf{u})$  для удаления нейрона и слоя.

### Теорема (Грабовой, 2021)

Пусть задано распределения вектора параметров  $p(\mathbf{u})$ . Тогда распределение вектора параметров  $p(\mathbf{v}) = \int_{\mathbf{v}_2 \in \mathbb{R}^{n_t-1}} p(\bar{\mathbf{v}}_1 | \mathcal{D}, \mathbf{v}_1 = \mathbf{0}) d\mathbf{v}_2$ , где  $\mathbf{v} = \phi_{RNN}(t, \mathbf{u})$ .

### Теорема (Грабовой, 2021)

Пусть апостериорное распределение параметров  $p(\mathbf{u} | \mathcal{D}) = \mathcal{N}(\mathbf{m}, \Sigma)$ , число слоев модели учителя равняется числу слоев модели ученика  $T = L$ , а размеры соответствующих пространств не совпадают  $n_t \geq n_t$ . Тогда распределение параметров  $p(\mathbf{v} | \mathcal{D})$ , где  $\mathbf{v} = \phi_{RNN}(t, \mathbf{u})$  также является нормальным.

### Теорема (Грабовой, 2021)

Пусть апостериорное распределение параметров  $p(\mathbf{u} | \mathcal{D}) = \mathcal{N}(\mathbf{m}, \Sigma)$ , соответствующие размеры пространств совпадают,  $n_t = n_{t+1}$ . а функция активации удовлетворяет свойству идемпотентности  $\sigma \circ \sigma = \sigma$ . Тогда апостериорное распределение задается плотностью:

$$p(\psi_{RNN}(t, \mathbf{u}) | \mathcal{D}) = \mathcal{N}(\mathbf{m}_v + \Sigma_{v, \bar{v}} \Sigma_{\bar{v}, \bar{v}}^{-1} (\mathbf{i} - \bar{v}), \Sigma_{v, v} - \Sigma_{v, \bar{v}} \Sigma_{\bar{v}, \bar{v}}^{-1} \Sigma_{v, \bar{v}}).$$

## Последовательность выравнивающих преобразований

Множество всех структур задается последовательностью натуральных чисел:

$$\mathfrak{H} = \{(n_1, n_2, \dots, n_T), \quad n_i \in \mathbb{N}, T \in \mathbb{N}\}.$$

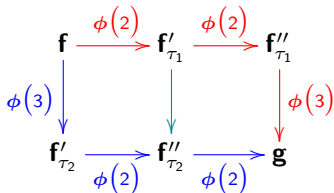
Множество структур, порождаемое структурой  $(n'_1, n'_2, \dots, n'_L)$  учителя  $\mathbf{f}$ :

$$\mathfrak{H}_{\mathbf{f}} = \{(n_1, n_2, \dots, n_T); \quad n_i \in \mathbb{N}, L \in \mathbb{N}; \quad n_i \leq n'_i, 3 \leq T \leq L; \quad n_{T-1} \leq n'_i, i > T\}$$

описывает конечное подмножество структур в континуальном множестве.

### Теорема (Грабовой, 2021)

Для произвольной структуры  $s \in \mathfrak{H}_{\mathbf{f}}$  существует последовательность локальных выравнивающих преобразований  $\tau = (\dots, \phi, \dots, \psi, \dots)$  сохраняющее распределение параметров модели  $\mathbf{f}$ .



**Пример:** Учитель  $\mathbf{f}$  имеет структуру  $(10, 6, 7, 10)$ , модель ученика  $\mathbf{g}$  имеет структуру  $(10, 4, 6, 10)$ .

**Последовательности выравнивающих преобразований  $\mathbf{f}$  в  $\mathbf{g}$ :**

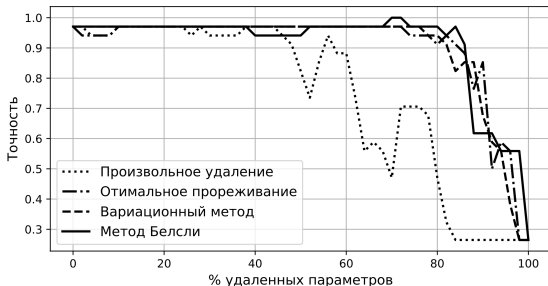
$$\tau_1 = (\phi(2), \phi(2), \phi(3)), \tau_2 = (\phi(3), \phi(2), \phi(2)), \tau_3 = (\phi(2), \phi(3), \phi(2))$$

Последовательность выравнивающих преобразований существует, но не единственно.

# Введение отношения порядка на множестве параметров

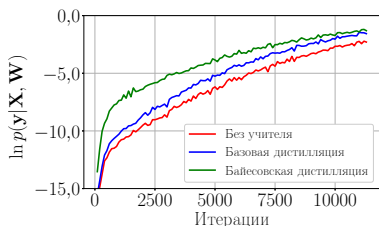
Порядок на множестве параметров задается:

- 1) случайным образом (базовая гипотеза),
- 2) оптимальным прореживанием:  $\xi = \arg \min_j h_{jj} \frac{u_j^2}{2}$ , где  $h_{jj}$  коэффициент при квадратичном члене в разложении функции ошибки  $\mathcal{L}$  по параметрам  $\mathbf{u}$ ,
- 3) отношением плотности апостериорного распределения параметра в нуле к значению параметра:  $\xi = \arg \max_j \frac{p(0|\mathbf{X},\mathbf{t})}{p(u_j|\mathbf{X},\mathbf{t})}$ ,
- 4) методом Белсли:  
 $\xi = \arg \max_j \frac{\lambda_{\max}}{\lambda_j}$ , где  $\lambda_j$  — сингулярные числа ковариационной матрицы параметров,
- 5) ковариационной матрицей градиентов функции ошибки  $\mathcal{L}$  по параметрам  $\mathbf{u}$ .





# Анализ вероятностных свойств ответов модели ученика



Модель учителя:

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

Модель ученика:

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

	Учитель	Без учителя	Баз. дист.	Байес. дист.
Структура	[10, 100, 50, 1]	[10, 50, 1]	[10, 50, 1]	[10, 50, 1]
Числ. парам.	6050	550	550	550
Инт. метр.	-	0	10244	25506

Правдоподобие модели ученика на основе **байесовской дистилляции** растет быстрее чем правдоподобие модели ученика на основе **базовой дистилляции Хинтона**.

Выборка	Модель	Кросс-энтропийная ошибка	Кросс-энтропийная ошибка с истинными вероятностями	Вероятностная разность	Точность	Число Параметров
FashionMnist	с учителем	$0,453 \pm 0,003$	-	$0,84 \pm 0,13$	$0,842 \pm 0,002$	7850
	без учителя	$0,461 \pm 0,005$	-	$0,86 \pm 0,18$	$0,841 \pm 0,002$	7850
Synthetic	с учителем	$0,618 \pm 0,001$	<b><math>1,17 \pm 0,05</math></b>	<b><math>0,45 \pm 0,20</math></b>	$0,828 \pm 0,002$	33
	без учителя	$0,422 \pm 0,002$	$2,64 \pm 0,02$	$0,75 \pm 0,22$	$0,831 \pm 0,001$	33
Twitter Sent.	с учителем	$0,489 \pm 0,003$	-	$0,79 \pm 0,17$	$0,764 \pm 0,005$	1538
	без учителя	$0,501 \pm 0,006$	-	$0,83 \pm 0,22$	$0,747 \pm 0,004$	1538

Модель с учителем аппроксимирует истинные вероятности классов.

Модель с учителем имеет меньшую разность между вероятностями классов.

## Выносятся на защиту

1. Предложен байесовский метод выбора моделей с использованием модели учителя с привилегированной и накопленной информацией.
2. Доказаны теоремы о свойствах дистилляции,
  - теоремы об эквивалентности для дистилляции моделей в случае задачи регрессии и классификации,
  - теоремы о виде априорного распределения параметров модели ученика в байесовской дистилляции.
3. Предложен метод выравнивания структур параметрических моделей. Предложен метод выбора априорного распределения параметров модели ученика с использованием апостериорного распределения параметров модели учителя для случаев
  - различных размерностей пространств параметров отдельных слоев,
  - различного в числа слоев нескольких моделей.
4. Предложены методы задания порядка на множестве параметров моделей
  - на основе корреляции параметров,
  - на основе оценки скорости сходимости параметров.
5. Предложена вероятностная интерпретации дистилляции моделей глубокого обучения. Исследованы свойства дистилляции моделей глубокого обучения.

# Список работ автора по теме диссертации

## Публикации в журналах ВАК

1. *Грабовой А.В., Стрижов В.В.* Байесовская дистилляция моделей глубокого обучения // Автоматика и телемеханика, 2021.
2. *Грабовой А.В., Стрижов В.В.* Анализ выбора априорного распределения для смеси экспертов // Журнал вычислительной математики и математической физики, 2021.
3. *A. Grabovoy, V. Strijov.* Quasi-periodic time series clustering for human // Lobachevskii Journal of Mathematics, 2020.
4. *Грабовой А.В., Бахтеев О. Ю., Стрижов В.В.* Введение отношения порядка на множестве параметров аппроксимирующих моделей // Информатика и ее применения, 2020.
5. *Грабовой А.В., Бахтеев О.Ю., Стрижов В.В.* Определение релевантности параметров нейросети // Информатика и ее применения, 2019.
6. *Грабовой А.В., Стрижов В.В.* Вероятностная интерпретация задачи дистилляции // Автоматика и телемеханика, 2022.

## Выступления с докладом

1. Задача обучения с экспертом для построения интерпретируемых моделей машинного обучения, Международная конференция «Интеллектуализация обработки информации», 2020.
2. Привилегированная информация и дистилляция моделей, Всероссийская конференция «63-я научная конференция МФТИ», 2020.
3. Введение отношения порядка на множестве параметров нейронной сети, Всероссийская конференция «Математические методы распознавания образов ММРО», 2019.
4. Анализ априорных распределений в задаче смеси экспертов, Всероссийская конференция «62-я научная конференция МФТИ», 2019.
5. Поиск оптимальной модели при помощи алгоритмов прореживания, Всероссийская конференция «61-я научная конференция МФТИ», 2018.
6. Автоматическое определение релевантности параметров нейросети, Международная конференция «Интеллектуализация обработки информации», 2018.