

# Определение параметров нейросети, подлежащих оптимизации

Кирилл Федорович Муравьев

Московский физико-технический институт

*Курс: Численные методы обучения по прецедентам  
(практика, В. В. Стрижов)/Группа 594, весна 2018*

# Цель исследования

## Цели исследования

Получить метод определения параметров, не требующих дальнейшей оптимизации.

## Проблемы

- Обучение глубоких нейросетей требует больших вычислительных ресурсов.
- Разные параметры нейросети сходятся с разной скоростью.

## Было предложено

- Определять параметры, дальнейшая оптимизация которых не принесет результата, и удалять их из множества оптимизируемых параметров.
- Определять параметры, влияние которых на предсказания сети мало, и удалять их из множества весов сети.

- Alex Graves: Practical Variational Inference for Neural Networks - вероятностный подход к оптимизации нейросетей
- Max Welling, Yee Whye Teh: Bayesian Learning via Stochastic Gradient Langevin Dynamics - оптимизация с помощью стохастической динамики
- Chunyan Li, Changyou Chen, David Carlson, Lawrence Carin: Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks. 2015 - SGLD с матрицей коэффициентов вместо константного шага

- $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$  - выборка.
- $f_{\mathbf{w}_0} : \mathbb{R}^n \rightarrow \mathbb{Y}$  - модель.
- $T$  - градиентный оператор:  $T(\mathbf{w}) = \mathbf{w} - \gamma \nabla L(\mathbf{w}, \mathcal{D}')$ .
- $\alpha$  - вектор оптимизируемых параметров,  $\beta$  - вектор параметров, включенных в нейросеть.
- $T|_{\alpha}(\mathbf{w}) = \mathbf{w} - \gamma \alpha \odot \nabla L(\mathbf{w}, \mathcal{D})$  - градиентный оператор для вектора  $\alpha$  параметров, включенных в оптимизацию.
- Ищем оптимальные вектора  $\alpha$  и  $\beta$ :

$$\arg \min_{\alpha \in \{0,1\}^N, \|\alpha\|_1=k} \min_{\beta \in \{0,1\}^N, \|\beta\|_1=l} L(\beta \odot (T|_{\alpha})^{\eta_1}(T^{\eta_0}(\mathbf{w}_0)), \mathcal{D})$$

**Подлежат оптимизации:** параметры с максимальной суммой абсолютных значений градиента функции потерь за последние  $t$  шагов:

$$\alpha = \arg \max_{\alpha \in \{0,1\}^N, \|\alpha\|_1 = k} \alpha \cdot \sum_{i=0}^{t-1} |\nabla \mathcal{L}(T^{\eta_0-i}(\mathbf{w}_0), \mathcal{D})|$$

**Остаются в сети:** параметры с наибольшими абсолютными значениями весов, веса остальных занулим:

$$\beta = \arg \max_{\beta \in \{0,1\}^N, \|\beta\|_1 = l} \beta \cdot |(T|_{\alpha})^{\eta_1}(T^{\eta_0}(\mathbf{w}_0))|$$

# Выбор параметров по ковариационной матрице

$C$  - ковариационная матрица стохастического градиента:

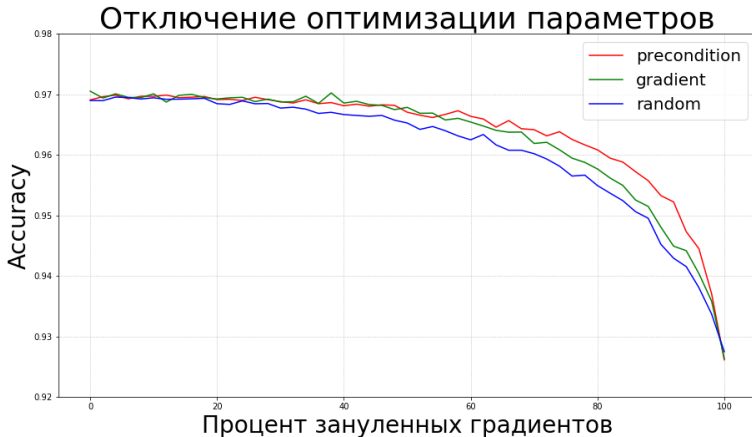
$$\nabla L(\mathbf{w}, \mathcal{D}') \sim \mathcal{N}(g, C)$$

$H$  - *precondition*-матрица:

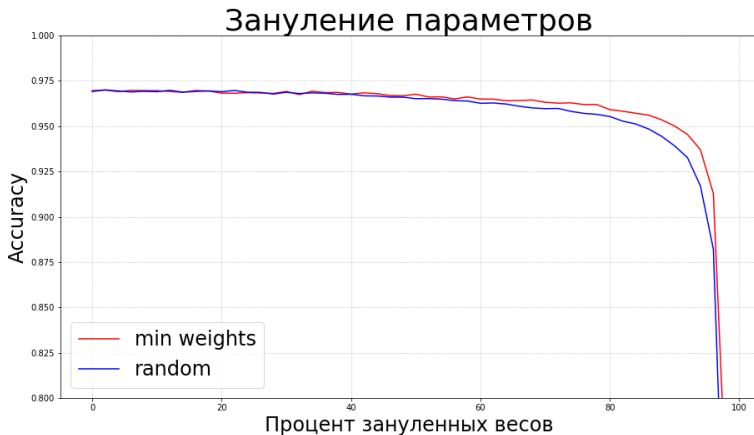
$$H = \frac{2S}{NC}; T(\mathbf{w}) = \mathbf{w} - H \nabla L(\mathbf{w}, \mathcal{D}')$$

**Подлежат оптимизации:** параметры с максимальными произведениями *precondition* и градиента:

$$\alpha = \arg \max_{\alpha \in \{0,1\}^N, \|\alpha\|_1=k} \alpha \odot |H \cdot \nabla \mathcal{L}(T^{\eta_0}(\mathbf{w}_0), \mathcal{D})|$$



Зависимость качества от доли параметров с отключенной оптимизацией



Зависимость качества от доли зануленных весов



- Представлена формальная постановка задачи для выборочной оптимизации глубоких нейросетей.
- Построена экспериментальная зависимость качества предсказания от доли оптимизируемых параметров и от степени прореживания нейросети.
- Проведено сравнение разных методов выборочной оптимизации.