УДК 519.853.62:004.032.26

СПЕКТРАЛЬНЫЙ АНАЛИЗ ДИНАМИКИ ОБУЧЕНИЯ НЕЙРОННЫХ СЕТЕЙ: ИССЛЕДОВАНИЕ НИЗКОРАНГОВОЙ СТРУКТУРЫ ГРАДИЕНТНОГО СПУСКА

Венско В.В., Якуть А.Ю., студенты

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Примичева З.Н. – канд. физ.- мат. наук, доцент

Аннотация. Данная работа посвящена исследованию структуры матриц весов, возникающей в процессе обучения многослойного перцептрона. Выдвинута гипотеза об убывании эффективного ранга матриц весов в процессе обучения методом градиентного спуска. Для проверки гипотезы обучена полносвязная нейронная сеть (MLP) на наборе данных MNIST с помощью библиотеки PyTorch; на каждой эпохе вычислялся эффективный ранг матриц весов для разных функций активации. Полученные результаты подтверждают гипотезу и демонстрируют устойчивую тенденцию к снижению эффективного ранга.

Ключевые слова. Нейронные сети, сингулярное разложение, эффективный ранг, энтропия Шеннона, градиентный спуск, MLP, ReLU, GELU.

Современные нейронные сети содержат миллиарды и триллионы параметров, однако ряд исследований указывает на то, что обучение фактически происходит в пространстве значительно меньшей размерности. Данное явление, называемое низкоранговой динамикой обучения, имеет важные практические следствия, поскольку может частично объяснять эффективность низкоранговых методов параметризации и адаптации, таких как LoRA (Low-Rank Adaptation).

В работе Xu et. al. [1, с. 6] показано, что при обучении двухслойной нейросети с гладкими функциями активации изменения весов фактически происходят в пространстве небольшой размерности – порядка $2K$, где K – число классов. Однако данные данные факты верны при определённых условиях инициализации и процессе обучения.

Целью данной работы – экспериментально исследовать динамику эффективного ранга матриц весов при обучении MLP в условиях, когда предпосылки описываемые в работе Xu et. al. [1] не выполнены: нарушены гладкость активации, двухслойность архитектуры, вид функции потерь и тип алгоритма оптимизации. Математическим аппаратом исследования служат сингулярное разложение матриц, понятие эффективного ранга и энтропия Шеннона.

Сравнивается динамика обучения при использовании недифференцируемой функции активации ReLU с гладкими дифференцируемыми функциями активации. Математическими инструментами исследования являются метод сингулярного разложения матриц, теория ранга и энтропия информационных систем.

Выдвигается следующая гипотеза: при обучении многослойного перцептрона на наборе данных MNIST эффективный ранг весовых матриц имеет тенденцию к снижению по мере обучения, причём наиболее заметно данное снижение проявляется в ранних слоях сети. Для проверки гипотезы будем использовать понятие эффективного ранга матриц [2].

Математический аппарат и постановка задачи. Пусть $W \in \mathbb{R}^{m \times n}$ – матрица весов линейного слоя нейронной сети в момент времени (эпохи) t. Матрица весов W задаёт линейный оператор

$$W\colon \mathbb{R}^{n} \to \mathbb{R}^{m}.$$

Размерность образа этого оператора равна рангу матрицы W, то есть

$$\dim(\mathrm{Im}W) = \mathrm{rank}(W).$$

Для анализа спектральной структуры матрицы весов используется сингулярное разложение (SVD – singular value decomposition).

Теорема (о сингулярном разложении) [3]. Для любой вещественной матрицы $A \in \mathbb{R}^{m \times n}$ существует разложение

$$A = U\Sigma V^{T},$$

где $U \in \mathbb{R}^{m \times m}$ и $V \in \mathbb{R}^{n \times n}$ – ортогональные матрицы; $\Sigma \in \mathbb{R}^{m \times n}$ – прямоугольная диагональная матрица с элементами $\sigma_1 \geq \cdots \geq \sigma_i \geq \cdots \geq \sigma_r \geq 0$ на главной диагонали, где $r = \mathrm{rank}(A)$. Числа σ_i

называются сингулярными числами матрицы A, столбцы U – левыми, столбцы V – правыми сингулярными векторами.

Геометрически сингулярное разложение означает, что любой линейный оператор, задаваемый матрицей W, допускает разложение в три последовательных действия: поворот в исходном пространстве, масштабирование, поворот в целевом пространстве.

Матрицу W можно записать в виде суммы матриц ранга 1 [3]

$$W = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T .$$

Каждое слагаемое $\sigma_i u_i v_i^T$ представляет собой матрицу единичного ранга, вклад которой в суммарное преобразование пропорционален числу σ_i. Если первые k сингулярных чисел значительно превосходят остальные сингулярные числа $\sigma_{k+1} \approx 0$, то матрица W хорошо аппроксимируется усечённым разложением вида [3]:

$$W \approx W_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T , \quad \operatorname{rank}(W_k) = k \ll \operatorname{rank}(W).$$

Ранг матрицы является дискретной характеристикой, потому что он может быть равен только целым неотрицательным числам. В реальных матрицах весов нейронных сетей, возникающих в результате обучения, все сингулярные числа ненулевые. Однако часть из них зачастую пренебрежимо мала по сравнению с другими. Существует мера, оценивающая не только факт неравенства нулю сингулярных чисел, но их относительную величину. Для количественной оценки концентрации сингулярного спектра используется понятие эффективного ранга, введённое Роем и Веттерли [2]. Пусть $\sigma_1 \geq \cdots \geq \sigma_i \geq \cdots \geq \sigma_r > 0$ – ненулевые сингулярные числа матрицы W. Определим нормированное распределение вероятностей:

$$p_i = \frac{\sigma_i}{\sum_{j=1}^{r} \sigma_j}, \quad i = 1, \ldots, r.$$

Обозначим через $\operatorname{erank}(W)$ эффективный ранг матрицы W. Тогда эффективный ранг матрицы W определяется как:

$$\operatorname{erank}(W) = e^{\left(-\sum_{j=1}^{r} p_i \ln p_i\right)} = e^{H(p)},$$

где $H(p)$ – энтропия Шеннона [4, с. 80] нормированного спектрального распределения. Величина $\operatorname{erank}(W)$ принимает значения $[1, r]$: значение $\operatorname{erank}(W) = r$ соответствует равномерному распределению сингулярных чисел, значение $\operatorname{erank}(W) \approx 1$ – случаю, когда единственное сингулярное число доминирует над остальными.

Рассмотрим двухслойную нейронную сеть:

$$f(X) = W_2 \varphi(W_1 X), \quad W_1 \in \mathbb{R}^{m \times d}, \quad W_2 \in \mathbb{R}^{K \times m},$$

где $X \in \mathbb{R}^{d \times N}$ – матрица входных данных (d – мерные векторы, N объектов), $Y \in \mathbb{R}^{K \times m}$ – матрица меток, K – число классов, φ – функция активации. В процессе обучения подбираются матрицы весов W_1, W_2, минимизирующие функцию потерь (например, среднеквадратичную ошибку):

$$\min_{W_1} \mathcal{L}(W_1) = \frac{1}{2} \| f_{W_1}(X) - Y \|_F^2 .$$

Вычислим градиент $\mathcal{L}$ по W_1 методом обратного распространения ошибки (backpropagation). Обозначим:

$$\Delta_2 = W_2 \varphi(W_1 X) - Y \in \mathbb{R}^{K \times N},$$

$$\Delta_1 = W_2^T \Delta_2 \odot \varphi'(W_1 X) \in \mathbb{R}^{m \times N},$$

где $\odot$ означает поэлементное произведение, а φ' – производную функции активации. Тогда

$$G_1 = \nabla_{W_1}\mathcal{L} = \Delta_1 X^T = \left(W_2^T \Delta_2 \odot \varphi'(W_1 X)\right) X^T.$$

Обучение ведётся градиентным спуском по функции потерь по формуле:

$$W_1(t+1) = W_1(t) - \eta \cdot G_1(t).$$

Методология эксперимента.

Обучение нейросети производится на наборе данных MNIST. MNIST содержит 70000 размеченных изображений рукописных цифр (60000 обучающих и 10000 тестовых), каждое размером 28 × 28 пикселей в градациях серого (рис. 1).

Рисунок 1 – Примеры изображений из набора данных MNIST

Каждое изображение разворачивалось в вектор $x \in \mathbb{R}^{784}$, нормирующийся по формуле:

$$X_{\text{norm}} = \frac{x - \mu}{\sigma}, \quad \mu = 0.1307, \quad \sigma = 0.3081,$$

при этом используется нейронная сеть (многослойный перцептрон) архитектуры $784 \to 512 \to 256 \to 128 \to 10$ с функцией активации ReLU между скрытыми слоями. Каждый линейный слой задается матрицей весов вида:

$$W^{(1)} \in \mathbb{R}^{512\times 784}, \quad W^{(2)} \in \mathbb{R}^{256\times 512}, \quad W^{(3)} \in \mathbb{R}^{128\times 256}, \quad W^{(4)} \in \mathbb{R}^{10\times 128}.$$

Отсюда ранги матриц удовлетворяют условиям

$$\text{rank}\left(W^{(1)}\right) \le 512, \quad \text{rank}\left(W^{(2)}\right) \le 256, \quad \text{rank}\left(W^{(3)}\right) \le 128, \quad \text{rank}\left(W^{(4)}\right) \le 10.$$

Такие же ограничения области значений будет иметь и эффективный ранг.

Обучение велось оптимизатором Adam с шагом $\eta = 0.001$ с функцией потерь CrossEntropyLoss в течение 50 эпох, размер мини-батча – 256. Зафиксировано начальное зерно генератора случайных чисел со значением 42. На каждой эпохе t для каждой весовой матрицы W выполнялись следующие шаги:

Шаг 1. Вычисление сингулярных чисел: $\sigma_i = \text{svdvals}(W) \in \mathbb{R}^n$, $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_n \ge 0$.

Шаг 2. Фильтрация: удаление $\sigma_i < 10^{-10}$ во избежание численных погрешностей при вычислении логарифма.

Шаг 3. Нормировка:

$$p_i = \sigma_i / \sum_j^n \sigma_j.$$

Шаг 4. Расчёт энтропии:

$$H = -\sum_i^n p_i \ln p_i.$$

Шаг 5. Оценка эффективного ранга:

$$\mathrm{erank}(W) = exp(H).$$

Реализация эксперимента производилась на языке программирования Python с использованием библиотеки PyTorch. Функция библиотеки PyTorch linalg.svdvals вычисляет сингулярные числа без построения полного SVD-разложения, что обеспечивает эффективность вычислений.

Результаты эксперимента.

На рисунке 2 представлена динамика эффективного ранга ($\mathrm{erank}(W(t))$) для всех четырёх слоёв в зависимости от номера эпохи $t = 1, \ldots, 50$. Наблюдается тенденция к убыванию эффективного ранга во всех слоях, что подтверждает гипотезу. За 50 эпох обучения достигается точность классификации 98,4% на тестовой выборке, модель достигает сходимости.

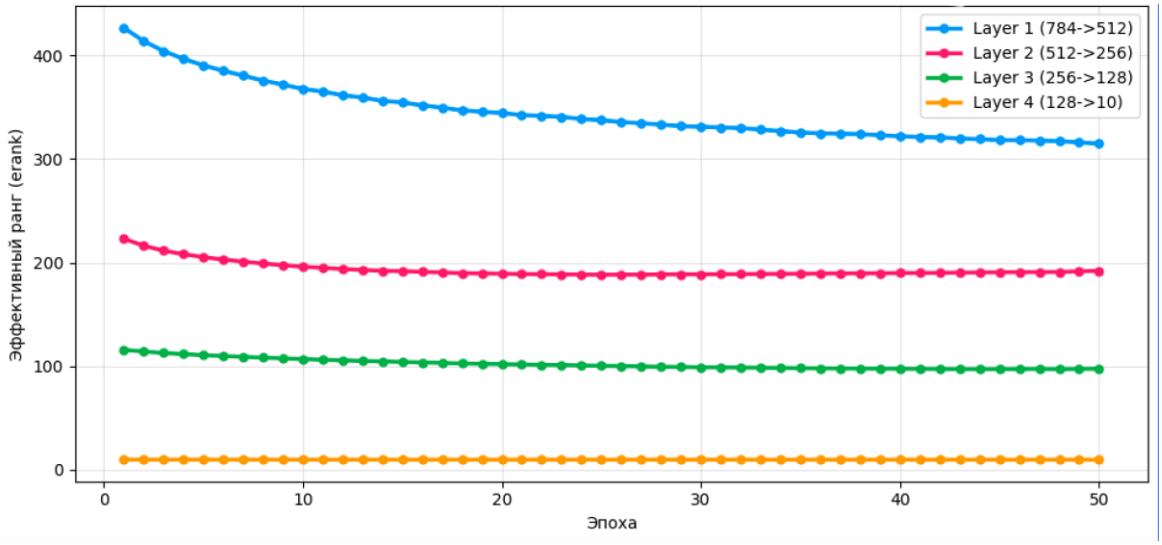

Рисунок 2 – Динамика эффективного ранга весовых матриц в процессе обучения с разными функциями активации.

В таблице 1 приведены значения эффективного ранга, отношения эффективного ранга к максимальному возможному после 50 эпох обучения.

Таблица 1 – Эффективный ранг матриц после 50 эпох обучения

Слой	макс. Ранг	Erank	erank/max_rank
$784 \rightarrow 512$	512	314.7	61.46
$512 \rightarrow 256$	256	192.7	75.27
$256 \rightarrow 128$	128	97.3	76.01
$128 \rightarrow 10$	10	9.8	98.00

В матрицах весов, особенно ранних слоёв, выделяется малое количество значимых сингулярных чисел, тогда как остальные стремятся к нулю. На первом слое (784→512) максимально возможный ранг равен 512, а эффективный ранг составил 314.7 – отношение $\mathrm{erank}(W^{(1)})/\mathrm{max\_rank}(W^{(1)}) = 0.615$ указывает на значительную неравномерность сингулярного спектра. Следует подчеркнуть, что это отношение характеризует концентрацию сингулярных чисел.

Для последнего слоя $(128 \rightarrow 10)$ $\mathrm{erank}\left(W^{(4)}\right) = 9.8 \approx 10$ объясняется архитектурным ограничением: $\mathrm{rank}\left(W^{(4)}\right) \leq \min(128, 10) = 10$, а теоретический максимум $\mathrm{erank}\left(W^{(4)}\right)$ при этом ранге равен ровно 10. Поскольку последний слой непосредственно осуществляет разделение объектов на 10 классов, высокая величина эффективного ранга в этом случае не противоречит общей гипотезе. Напротив, она показывает, что снижение эффективного ранга выражено неодинаково в разных частях сети и сильнее проявляется в ранних слоях.

На рисунке 3 представлена тепловая карта нормированных сингулярных чисел. Характерный «обрыв» в распределении сингулярных чисел в ранних слоях указывает на сосредоточение преобразования в низкоразмерном подпространстве.

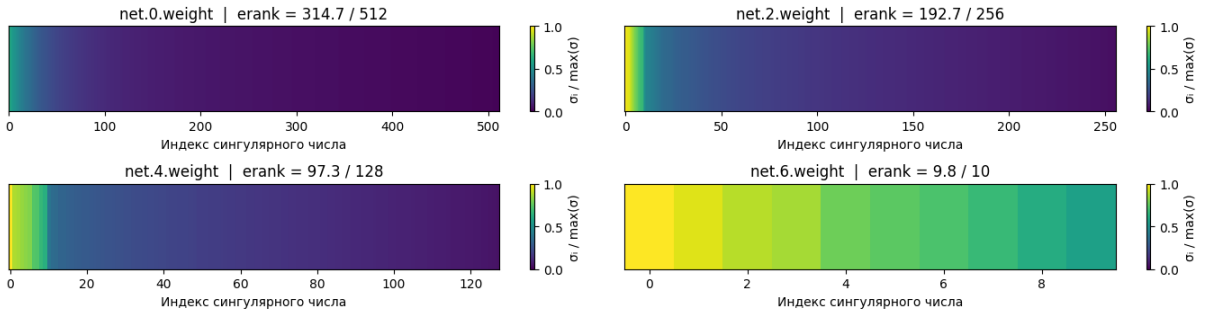

Рисунок 3 – Тепловая карта распределения сингулярных чисел

Обратим внимание, что в нашем эксперименте используется функция активации ReLU, формально не удовлетворяющая условию гладкости функции. Тем не менее, согласно Xu et. al. [1], явление низкоранговой динамики приближенно сохраняется и для негладких активаций. Поскольку исходная теория обновления весов [1] связана с гладкими функциями, сравним эксперимент с другими функциями активации библиотеки PyTorch: tanh, sigmoid, GELU.

При обучении MLP с использованием гладких функций активации (гиперболический тангенс, сигмоида, GELU) наблюдается такое же явление – эффективный ранг имеет тенденцию к снижению. В силу задания функций активации ReLU и GELU эффективный ранг изменяется по схожей траектории. Можно построить графики и заметить, что они неограниченно близко приближаются к друг другу при аргументе стремящемся к бесконечности. Динамика эффективного ранга первого слоя для всех четырёх активаций показана на рисунке 4.

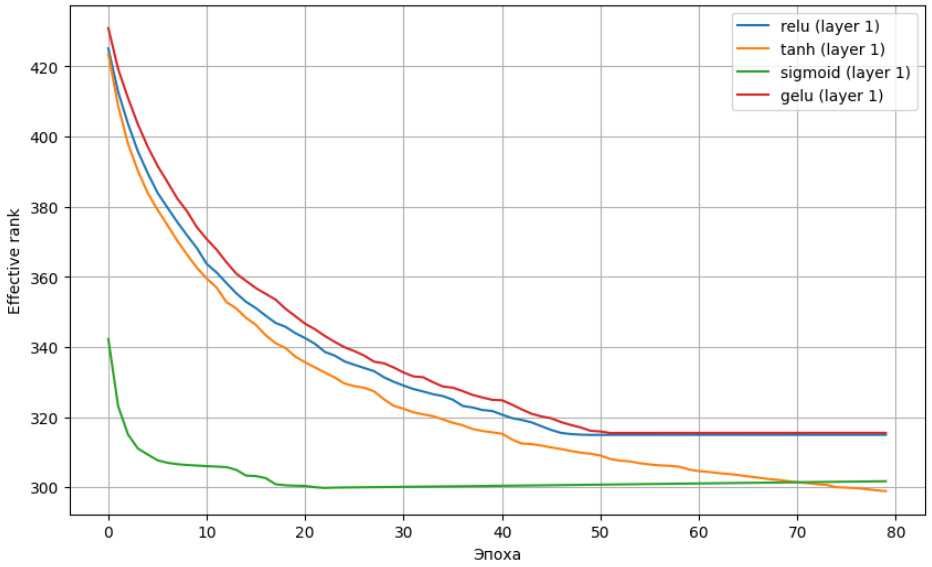

Рисунок 4 – динамика снижения эффективного ранга матриц первого слоя разных функций активации

Таким образом, в данной работе проведено экспериментальное исследование низкоранговой структуры, возникающей в процессе обучения многослойного перцептрона (MLP). Выдвинутая гипотеза подтверждается экспериментально: эффективный ранг снижается во всех слоях сети в процессе обучения методом градиентного спуска. Наиболее выраженное снижение наблюдается в ранних слоях большой размерности: первый слой (784→512) задействует лишь 61.46% своего пространства.

Практическим следствием дальнейшего изучения темы ранговых свойств градиентного спуска и весовых матриц является эффективность низкоранговых методов инициализации таких, как LoRA (Low-Rank Adaptation). Поскольку обучение фактически происходит в низкоразмерном подпространстве [1], явное ограничение ранга при правильной инициализации позволяет достигать сопоставимого качества при существенно меньшем числе параметров.

Перспективные направления дальнейшей работы: прямое измерение $\mathrm{erank}(\mathrm{W}(\mathrm{T}) - \mathrm{W}(0))$, увеличение числа независимых запусков с различными seed, повторение эксперимента на других архитектурах (CNN, Transformer) и наборах данных (CIFAR-10, Fashion-MNIST).

Список использованных источников.

1. Emergent Low-Rank Training Dynamics in MLPs with Smooth Activations / A. S. Xu, C. Yaras, M. Asato [et al.]. – 2026. – URL: https://arxiv.org/abs/2602.06208 (date of access: 09.02.2026).

2. Roy, O. The effective rank: a measure of effective dimensionality / O. Roy, M. Vetterli // EUSIPCO 2007 : proc. of the 15th Europ. signal processing conf., Poznan, 3–7 Sept. 2007. – Poznan, 2007. – P. 606–610.

3. Воеводин, В. В. Матрицы и вычисления / В. В. Воеводин, Ю. А. Кузнецов. – М. : Наука, 1984. – 318 с.

4. Чумак, О. В. Энтропии и фракталы в анализе данных / О. В. Чумак. – М. ; Ижевск : НИЦ «Регулярная и хаотическая динамика» : Ин-т компьютер. исслед., 2011. – 164 с.

5. MNIST Digit Recognition Challenge // Kaggle. – URL: https://www.kaggle.com/competitions/exploretechla-ignite-digit-recognition-challenge (date of access: 09.02.2026).

UDC 519.853.62:004.032.26

SPECTRAL ANALYSIS OF NEURAL NETWORK TRAINING DYNAMICS: INVESTIGATING THE LOW-RANK STRUCTURE OF GRADIENT DESCENT

Vensko V.V. Yakut A.Y., students

Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus
Primicheva Z.N. – PhD in Physics and Mathematics

Annotation. This work is devoted to studying the structure of weight matrices arising during the training of a multilayer perceptron. A hypothesis is proposed that the effective rank of weight matrices decreases during training by gradient descent. To test the hypothesis, a fully connected neural network (MLP) was trained on the MNIST dataset using the PyTorch library; at each epoch, the effective rank of the weight matrices was computed for different activation functions. The obtained results confirm the hypothesis and demonstrate a steady trend toward a decrease in the effective rank.

Keywords. Neural networks, singular value decomposition, effective rank, Shannon entropy, gradient descent, MLP, ReLU, GELU.