

doi: 10.17586/2226-1494-2026-26-5-993-999

УДК 004.032.26

Поиск перспективных точек инициализации нейронных сетей

Виктор Николаевич Толстых^{1✉}, Владимир Николаевич Ершов²

^{1,2} Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 19000, Российская Федерация

¹ victor.n.tolstykh@gmail.com✉, <https://orcid.org/0009-0003-0202-5500>

² vyershov@guap.ru, <https://orcid.org/0000-0002-9830-2159>

Аннотация

Введение. Исследована проблема инициализации нейронной сети, заключающаяся в определении начальной точки в пространстве весовых коэффициентов для качественного градиентного спуска. Известно, что скорость обучения и результат обучения нейронной сети во многом зависят от ее инициализации. Проблема заключается в том, что в процессе выполнения алгоритм может попасть в аномальную зону локального минимума и надолго там застрять, иногда в самом начале обучения. В таких случаях рекомендуется начать спуск с новой инициализации. В работе предложено начинать обучение с точки на дне локального минимума, с которой, как показывают исследования, начинается спуск в новом направлении. **Метод.** Предлагаемый метод основан на том, что большинство аномалий являются критическими точками, имеющими направление на спуск. Нейросеть инициализируется каждый раз, когда алгоритм обучения сталкивается с аномалией, в которой значение функции ошибки возрастает. При этом производится определение новой инициализации в точке минимума функции ошибки на отрезке, соединяющем точку до аномалии и следующую точку за ней. Последовательные инициализации приводят к тому, что спуск происходит по оптимизированной траектории без значимых осцилляций. **Основные результаты.** Работоспособность метода показана на примере перцептрона с единственным скрытым слоем и одной аномалией. Доказано, что спуск из новой точки инициализации происходит быстрее; обнаружена возможность коррекции градиента без вычисления гессиана; сделан вывод о целесообразности многократной инициализации нейронной сети при попадании в зону очередной аномалии. Предлагается использовать способ коррекции алгоритма градиентного спуска с целью достижения дна аномалии в случаях, когда ошибка начинает расти. **Обсуждение.** Разработанный метод гарантирует более высокую вероятность попадания в область с низким значением функции ошибки, чем при случайном розыгрыше точки инициализации, а также выигрыш в сравнении с ограниченной машиной Больцмана по времени исполнения.

Ключевые слова

обучение нейронной сети, градиентный спуск, инициализация нейросети, локальный минимум, регуляризация, функция ошибки

Благодарности

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации, соглашение № FSRF-2023-0003, «Фундаментальные основы построения помехозащищенных систем космической и спутниковой связи, относительной навигации, технического зрения и аэрокосмического мониторинга».

Ссылка для цитирования: Толстых В.Н., Ершов В.Н. Поиск перспективных точек инициализации нейронных сетей // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 5. С. 993–999. doi: 10.17586/2226-1494-2026-26-5-993-999

Search for promising points of neural networks initialization

Victor N. Tolstykh¹, Vladimir N. Yershov²^{1,2} Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, 190000, Russian Federation¹ victor.n.tolstykh@gmail.com, <https://orcid.org/0009-0003-0202-5500>² vyershov@guap.ru, <https://orcid.org/0000-0002-9830-2159>**Abstract**

The paper examines the problem of initialization, which consists in determining the starting point in the space of the neural network weights for a qualitative gradient descent. It is known that the learning speed and learning result of a neural network mostly depends on its initialization. The problem is that in the process of execution, the algorithm can get into the anomalous zone of the local minimum and get stuck there for a long time, sometimes at the very beginning of training. In such cases, it is recommended to start the descent with a new initialization. The study proposes to start training from a point at the bottom of the local minimum, from which, as research shows, the descent in a new direction begins. The proposed method is based on the assumption that most of these anomalies are critical points that have a direction to descent. The neural network is initialized every time the learning algorithm encounters an anomaly in which the value of the error function increases. At the same time, a new initialization is determined at the minimum point of the error function in the segment connecting the point before the anomaly and the following one. Successive initializations result in the descent occurring along an optimized trajectory without significant oscillations. The effectiveness of the method is demonstrated by the example of a perceptron with a single hidden layer and one anomaly. It is proved that the descent from the new initialization point is faster; this demonstrates the possibility of gradient correction without calculating the Hessian; a conclusion is made about the expediency of multiple initializations of the neural network when it enters the zone of the next anomaly. It is proposed to use the method of correcting the gradient descent algorithm to reach the bottom of the anomaly in cases where the error begins to grow. The developed method gives a high probability of hitting a region with a low value of the error function than in the case of a random drawing of the initialization point, as well as a benefit in comparison with a Restricted Boltzmann's machine in terms of execution time.

Keywords

neural network training, gradient descent, neural network initialization, local minimum, error function.

Acknowledgements

The work was carried out with the financial support of the Ministry of Science and Higher Education of the Russian Federation, Agreement No. FSRF-2023-0003, "Fundamental principles for building interference-proof systems for space and satellite communications, relative navigation, technical vision and aerospace monitoring".

For citation: Tolstykh V.N., Yershov V.N. Search for promising points of neural networks initialization. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 5, pp. 993–999 (in Russian). doi: 10.17586/2226-1494-2026-26-5-993-999

Введение

Одной из известных проблем, с которой сталкиваются разработчики и пользователи нейронных сетей является колебательное поведение функции ошибки, часто возникающее в процессе обучения нейронной сети [1, 2]. Это явление замедляет обучение нейронной сети. Существуют эффективные методы борьбы, в которых либо изменяется правило вычисления следующего шага метода градиентного спуска, либо меняется геометрия функции ошибки, зависящая как от архитектуры нейросети, так и от обучающей выборки: методы Momentum, Нестерова (Nesterov Accelerated Gradient, NAG), RMSProp, AdaDelta, Adam и другие, адаптируют поведение градиента вблизи аномальных областей¹ [3], а метод Левенберга–Марквардта [4] использует диагональ матрицы Гессе (гессиана) для более точного определения шага. Общим у этой группы методов является управление градиентным спуском в местах критических точек функции ошибки. Сама геометрия функции ошибки при этом остается неизменной. Другая группа методов: стохастический градиентный спуск SGD, активатор ReLU и регуляризатор Dropout меняют

геометрию функции ошибки [2, 5]. Оптимизаторами эти методы не называются, но они выполняют ту же функцию — произвольное изменение функции ошибки может привести к исчезновению локальной аномалии. Выбор конкретного метода оптимизации зависит от требуемого баланса точности и скорости обучения. Также часто применяется комбинация методов в одном коде, к примеру — SGD, ReLU, Adam и Dropout². Другим радикальным способом избавления от попадания в локальные аномалии является выбор подходящей точки инициализации нейросети, которая в настоящее время выбирается случайным образом. Проблемой поиска оптимальной точки инициализации нейронной сети занимались многие исследователи и продолжают заниматься до сих пор [6–8]. Есть оригинальные решения, среди них методы Ксавьера, Хе [9] и ограниченная машина Больцмана [9–11], но в большинстве своем они затратны по ресурсам. Проще начать обучение нейросети с новой инициализации если обучение вошло в резонанс на первых шагах. Исследование поведения функции ошибки (также функции потерь), ее ландшафта и методов оптимизации, проводились и проводятся многими авторами, что отражено, в частности, в работах [12, 13].

¹ Optimizers [Электронный ресурс]. Режим доступа: <https://ml-cheatsheet.readthedocs.io/en/latest/optimizers.html> (дата обращения: 24.09.2026).

² Dropout и Batch normalization [Электронный ресурс]. Режим доступа: <https://habr.com/ru/companies/mvideo/articles/782360/> (дата обращения: 24.09.2026).

Цель работы состоит в разработке концепции поиска точек инициализации для беспроblemного градиентного спуска, опирающейся на анализ характера локальных аномалий и многомерной геометрии функции ошибки.

Описание предлагаемого метода

Геометрия функции ошибки. Многие методы оптимизации строятся на предположении, что обучение может застрять в «локальном минимуме» функции ошибки и из него как-то нужно выбираться. Как выглядят локальные минимумы обычно иллюстрируется на примерах одномерной кривой или двумерной поверхности. Расширение этих представлений на более многомерные зависимости не совсем корректно и может ввести в заблуждение, особенно в отношении локальных минимумов. В настоящей работе исследуется геометрия функции ошибки $s(\mathbf{w})$, как некоторой n -мерной гиперповерхности в пространстве размерности $n + 1$, зависящей от обучающей выборки и архитектуры нейросети

$$s(\mathbf{w}) = \frac{1}{K} \sum_{k=1}^K (y(\mathbf{w}, \mathbf{x}_k) - y_k)^2,$$

где $y(\mathbf{w}, \mathbf{x})$ — перцептрон с единственным нейроном на выходе; $\{\mathbf{x}_k, y_k\}$, $k = 1, \dots, K$ — обучающая выборка; $\mathbf{w}$ — набор весовых коэффициентов.

Обучение нейронной сети ведется в многомерном пространстве весовых коэффициентов оптимизацией по критерию ошибки. Но, в связи с тем, что окружающая нас многомерность трехмерная, то нам свойственно представлять ее как колышущуюся или холмистую поверхность в трехмерном пространстве. С холмов всегда есть спуск, а впадины могут составлять проблему, поскольку не все они достаточно глубокие, чтобы быть глобальным минимумом. Это представление породило методы оптимизации, позволяющие проскакивать через локальные минимумы. В целом оптимизация решает проблему, но такие представления о локальных аномалиях не совсем соответствуют тому, что показывают эксперименты на многомерных моделях.

Необходимым признаком локального минимума является равенство частных производных функции ошибки нулю (теорема Ферма). Точка, в которой градиент обращается в ноль, не обязательно является точкой минимума, поэтому ее принято называть критической или стационарной точкой. Достаточным признаком является строгая положительная определенность гессиана, при которой все его собственные числа (или главные миноры) строго положительны. Наиболее общим случаем для многомерных гиперповерхностей является критическая точка, у которой знаки собственных чисел гессиана распределены примерно поровну. Это не максимум и не минимум, а скорее аналог седловой точки у поверхности в трехмерном пространстве, и у таких точек всегда есть направление на спуск. Будем говорить, что этот спуск идет по желобу. В окрестности седловой точки алгоритм может застрять, только если он попадет строго в самую седловую точку. Но, поскольку множество критических точек имеет меру нуль, то такое событие представляется невероятным.

Отсюда вывод: основной источник резонансов (они же колебания, осцилляции или флуктуации) — не локальные минимумы. Спуск функции ошибки ведет не в саму критическую точку, а в ее окрестность и путь, по которому он продолжит движение, будет в какой-то мере напоминать ложбину, ущелье, желоб, в котором выделенное направление указывает на спуск, а по другим ортогональным направлениям частные производные будут оставаться с нулевыми значениями. Таким образом, на основании рассуждений и моделирования можно сделать вывод, что попадание алгоритма градиентного спуска в наклонный желоб типично, а в настоящий локальный минимум, в котором гессиан строго положителен — практически невероятно.

Исследование геометрии функции ошибки.

Нейроны скрытых слоев в перцептроне не имеют изначальной иерархии, они все находятся в равных условиях, поэтому перестановка нейронов скрытого слоя не меняет результат работы нейросети. Хотя, меняет их координаты в пространстве весовых коэффициентов. Так возникают симметрии, которые характерны для геометрии функции ошибки.

На рис. 1 изображен простой двухслойный перцептрон как одномерная функция $y = f(\mathbf{x})$ от многомерного аргумента $\mathbf{x}$. Предположим, что перцептрон обучен и все весовые коэффициенты зафиксировали свои значения. Теперь можно трансформировать обученный перцептрон таким образом, чтобы первый «красный» нейрон стал вторым, а второй «зеленый» первым. С точки зрения конечного результата, ничего в архитектуре перцептрона не изменилось, поскольку нейрон выходного слоя y получит те же значения, что и в начальном варианте. Из этого следует, что у каждого желоба имеется много идентичных копий, по числу, соответствующему с числом перестановок нейронов скрытого слоя. Попадая в желоб, алгоритм обучения нейронной сети начинает отражаться от его стенок, не достигая дна. Была смоделирована ситуация с желобом (рис. 2)

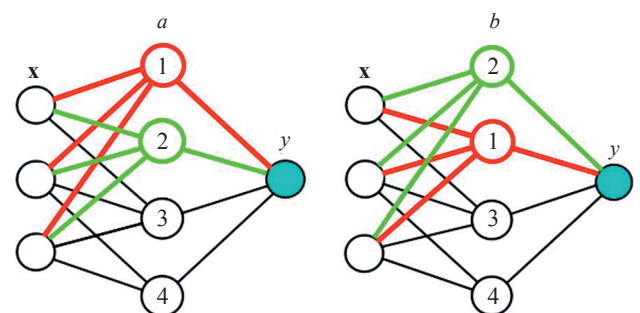


Рис. 1. Простой перцептрон: в оригинальном состоянии (a); в состоянии, при котором изменена нумерация нейронов — первый «красный» нейрон поменялся местами со вторым «зеленым» (b).

x — входящий вектор аргумента; y — значение на выходе перцептрона

Fig. 1. Simple perceptron: in (a) its original state and (b) in a state where the neuron numbering has been changed: the first “red” neuron has swapped places with the second “green” one. x is the incoming argument vector and y is the output value of the perceptron

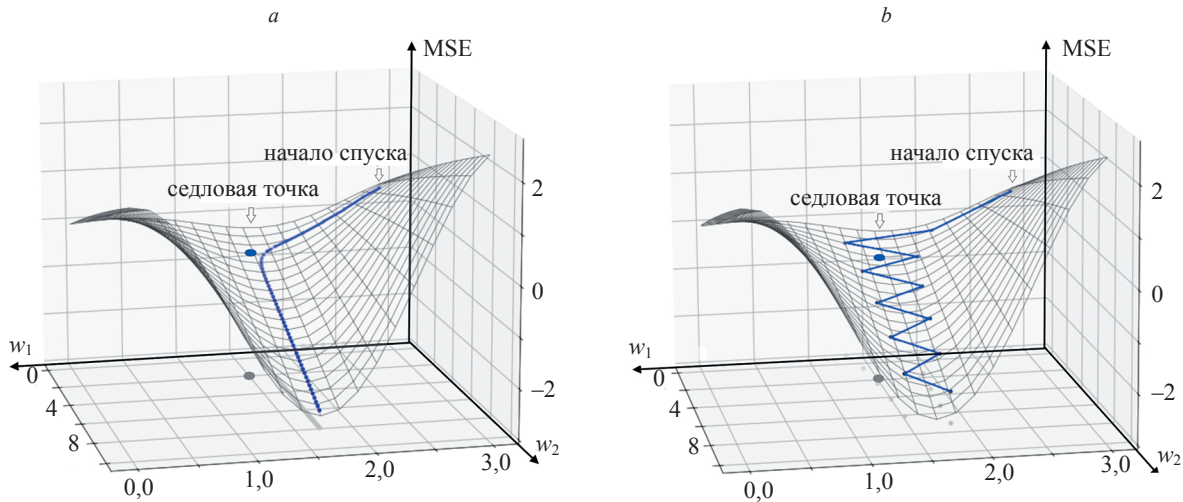


Рис. 2. Результат моделирования градиентного спуска функции ошибки $z = f(x, y)$ в двумерном пространстве аргумента $\mathbf{w}$: с малым шагом (а); с большим шагом, пропорциональным значению ошибки (б).

MSE — среднеквадратическая ошибка (Mean Squared Error)

Fig. 2. The result of modeling the gradient descent of the error function $z = f(x, y)$ in the two-dimensional space of the argument $\mathbf{w}$: with a small step (a), with a large step proportional to the error value (b)

на языке программирования Python + Matplotlib. В качестве окрестности критической точки выбрана сумма косинусоиды и параболы с аргументами по разным осям, имитирующими весовые коэффициенты w_1 и w_2 , и применен алгоритм градиентного спуска.

Модельная поверхность, содержащая седловую точку, собрана из суммы константы, смещенной косинусоиды по x и отрицательной параболы по y . Простейшая модель седловой точки — $z = x^2 - y^2$. В настоящей работе подобная модель, поскольку $\cos x \sim 1 + x^2$ в точке нуль.

На рис. 2, а показан вид градиентного спуска каким он должен был бы быть, а на рис. 2, б изображена реальная ситуация, в которой спуск осуществляется зигзагами и отражается от стенок желоба, не опускаясь на его дно. Результат моделирования доказал, что можно изменить градиент, укоротив его длину вдвое, когда он собирается «перепрыгнуть» через желоб и двигаться дальше. Тем самым, он попадет не на противоположную стенку, а опустится на дно желоба и в этой точке сменит направление спуска на более близкое к требуемому (рис. 2, а).

Корректировка градиента. На основе анализа модельной локальной аномалии, представленной многомерным желобом, а также изучении существующих алгоритмов оптимизации, использующих вычисление гессиана функции ошибки (алгоритм Левенберга–Марквардта), возникла идея помещать точку инициализации на дно желоба. Если значение функции ошибки в какой-то момент вдруг пошло вверх — явный признак пересечения желоба с резким изменением направления градиента. В этом случае он уже показывает на направление близкое к противоположному. Отрезок между двумя точками, на котором произошло резкое изменение градиента, либо приращение ошибки изменило знак, можно рассматривать как пространство одномерного аргумента функции ошибки с локальным

одномерным экстремумом внутри. Если это одномерный локальный минимум, то область значений аргумента функции ошибки можно задать на отрезке между точками аргумента $\mathbf{w}^{[k-1]}$ и $\mathbf{w}^{[k]}$

$$\mathbf{w}(t) = (1 - t)\mathbf{w}^{[k-1]} + t\mathbf{w}^{[k]}, t \in [0, 1].$$

Разбив единичный отрезок на n равных частей, получим дискретную зависимость $s(\mathbf{w}(t))$, где $t = [0, 1/n, 2/n, \dots, (n-1)/n, 1]$ от параметра t .

Локальный одномерный минимум найдем методом последовательных приближений. Чтобы оценить перспективность направления, достаточно взять среднюю точку, уменьшив шаг до половины длины градиента. Это довольно грубое приближение, но и в этом случае эксперимент позволяет понять, как ведет себя градиентный спуск после попадания на дно желоба.

Значение номера случайной последовательности в Python `seed = 4` на рис. 3 и 5 означает, что все эксперименты проводились в равных условиях.

Отметим, что изображение функции ошибки с единственной простой аномалией (рис. 3, а) подобно модели на рис. 2. Для чистоты эксперимента были подобраны параметры двухслойного перцептрона Розенблатта — соотношение числа синапсов входного слоя и числа нейронов скрытого слоя, а также подбор случайных чисел таким образом, чтобы аномалия была единственной. На рис. 3, а при первой встрече с аномалией производится корректировка — модуль градиента укорачивается вдвое, тем самым, шаг итерации попадает не на противоположную стенку, а примерно на дно желоба, и градиент меняет направление на идущее вдоль желоба. Некоторое число шагов градиент проходит в верном направлении, но затем снова упирается в противоположную стенку и требуется следующая корректировка (рис. 3, б). Причина изменения движения — в приближенности определения дна желоба.

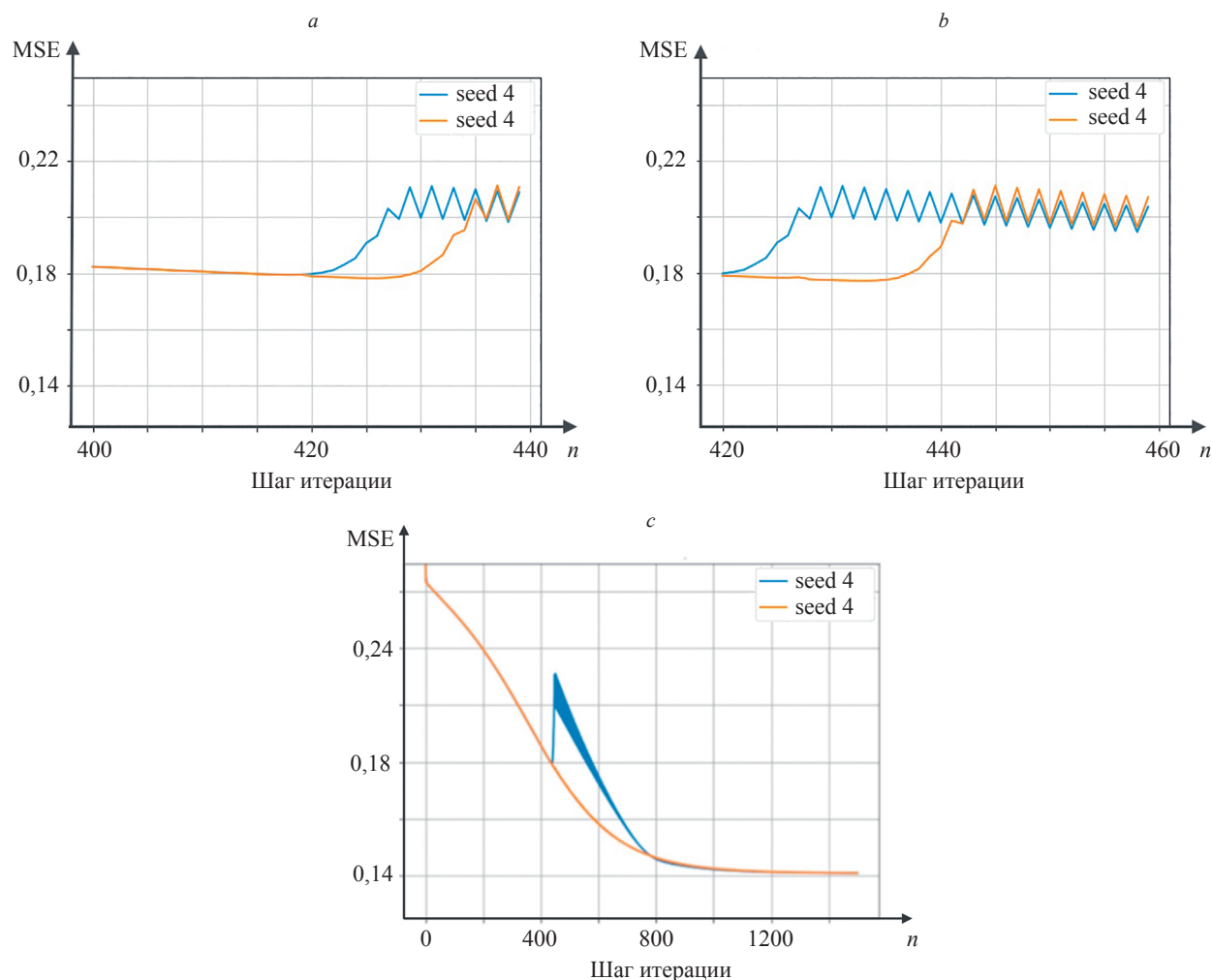


Рис. 3. Перцептрон с аномалией. Простая аномалия, представляющая типичный желоб; две корректировки градиента при попадании в аномалию (a, b); исходный и скорректированный путь (c) (алгоритм прошел по дну желоба с многочисленными корректировками)

Fig. 3. Perceptron with an anomaly. A simple anomaly representing a typical trough; two gradient adjustments when hitting an anomaly (a, b); the initial and corrected path — the algorithm passed along the bottom of the trough with numerous adjustments (c)

Результат последовательных корректировок показан на рис. 3, c и в первом приближении производит впечатление, что необходимый результат достигнут. Также можно предположить, что обучение нейросети возможно существенно ускорить если увеличить величину шага. Вместе с тем результаты моделирования на более сложных примерах показали, что данный алгоритм корректировки слишком прост, чтобы быть универсальным. Он не всегда сокращает осцилляцию ошибки, а в некоторых ситуациях даже может ее спровоцировать. Этот вопрос нуждается в дополнительных исследованиях.

Модельный пример

Был проведен численный эксперимент, позволивший выявить характер аномалий вида «локальный минимум». Исследовано поведение функции ошибки вдоль прямых, пересекающих пространство многомерного аргумента $\mathbf{w}$ в произвольно выбранных направлениях. В большинстве случаев получены спокойные участки, но также выявлены участки, которые содержали аномалии. Некоторые из них приведены на рис. 4.

Интервал поиска аномалии проводился на двухслойном перцептроне с использованием случайной точки $\mathbf{w}_0$ и случайного вектора (отрезка) направления $\mathbf{v}_0$, разбитого на $n = 200$ шагов. Выполнен ряд экспериментов с использованием серии параллельных линий (рис. 4, f), чтобы убедиться, что объектом исследования является желоб. В результате, рис. 4, f показал большое разнообразие аномалий. Первые две аномалии (рис. 4, a, b) имеют гладкий U-образный одномерный минимум, похожий на тот, который использовался при моделировании (рис. 2). Рис. 4, c имеет редко встречающийся W-образный двойной минимум, рис. 4, d, e — резкий V-образный минимум, случайное попадание в который исключено. Значения минимумов в серии с использованием параллельных линий не равны между собой, их изменение указывает на дно пологого желоба, на котором определяется следующая точка инициализации.

Спуск из точки инициализации нейронной сети.

Если аномалии действительно являются желобами то, можно предположить, что они указывают на более краткий путь к точке глобального минимума. Необходимо вовремя изменить направление градиента, при этом не

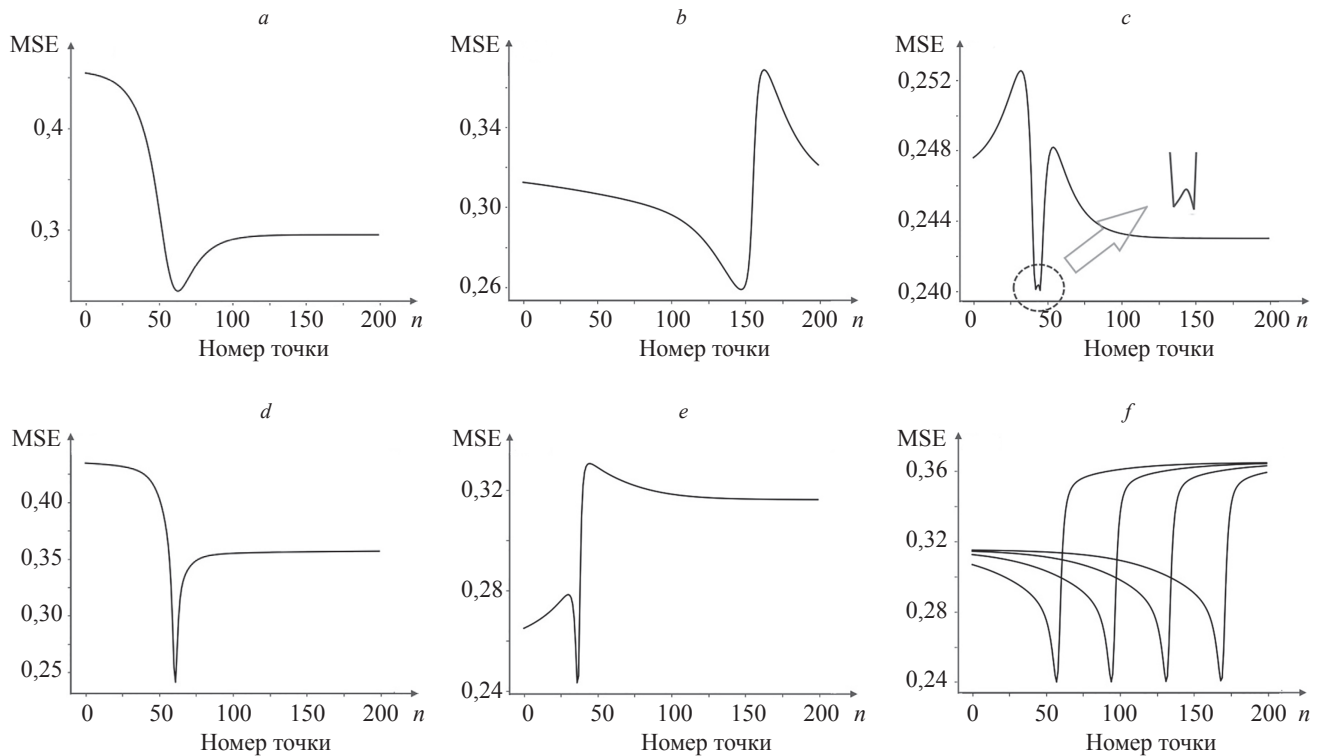


Рис. 4. Фрагменты поведения функции ошибки вдоль прямой в пространстве аргумента, в которых наблюдаются желоба: гладкие (a, b); двоянный (c); острые (d, e); узкий, который указывает на поведение функции вдоль четырех параллельных прямых (f)

Fig. 4. Fragments of the behavior of the error function along a straight line in the argument space, in which troughs are observed: smooth (a, b); double (c); sharp (d, e); narrow, showing the behavior of the function along four parallel lines (f)

пытаться аномалию проскочить или обойти, а опуститься на ее дно и пойти вдоль желоба.

На рис. 5 смоделирована ситуация, в которой из случайной точки пространства аргумента функции ошибки с начальной ошибкой 0,316 происходит гра-

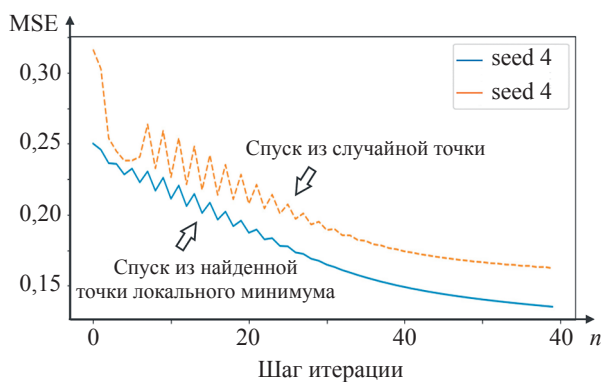


Рис. 5. Поведение градиентного спуска из случайной точки. Штриховая линия при обычном спуске, сплошная — при спуске из найденной точки инициализации

Fig. 5. Gradient descent behavior from a random point with error = 0.316 (stroke line) and from a point at the bottom of the trough with error = 0.250 (solid line)

диентный спуск (штриховая линия). Затем случайным перебором выполняется поиск прямой в пространстве аргумента функции ошибки, на которой определяется локальный минимум со значением ошибки 0,250 (точка инициализации нейросети), далее из полученной точки производится градиентный спуск без коррекции (сплошная линия). Видно, что в обоих случаях есть эффект резонанса, но во втором случае он имеет меньшую амплитуду, поскольку отражается от стенок вблизи дна желоба. Результаты моделирования частично приведены в работах [14, 15].

Заключение

Предлагаемый метод гарантирует более высокую вероятность попадания в область с низким значением функции ошибки, чем при случайном определении точки инициализации, а также в сравнении с ограниченной машиной Больцмана по времени исполнения. Он может быть использован совместно с методами контролируемой дисперсии Ксавьера и Хе. Отметим, что предложенный алгоритм корректировки схематичен и может быть реализован в разных вариантах, в зависимости от условий применения. Разработанный метод может стать полезным дополнением к существующим методам оптимизации.

Литература

1. Pascanu R., Mikolov T., Bengio Y. Understanding the exploding gradient problem // CoRR, 2012. V. 2. N 417. P. 1.
2. Николенько С., Кадури А., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей. СПб.: Питер, 2020. 480 с.
3. Li Y., Yuan Y. Convergence analysis of two-layer neural networks with ReLU activation // Proc. of the 31st Conference on Neural Information Processing Systems. 2017.
4. Gavin H.P. The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems // Department of Civil and Environmental Engineering Duke University. 2026. P. 1–22.
5. Julius, Mahale G., Sumana T., Adityakrishna C.S. On the modeling of error functions as high dimensional landscapes for weight initialization in learning networks // Proc. of the International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS). 2016. P. 202–210. doi: 10.1109/samos.2016.7818349
6. Bociort F., Serebriakov A., Braat J. Local optimization strategies to escape from poor local minima // International Optical Design Conference. 2002. P. ITuA5. doi: 10.1364/iode.2002.itua5
7. Rudd-Orthner R.N.M., Mihaylova L. Deep ConvNet: Non-random weight initialization for repeatable determinism, examined with FSGM // Sensors. 2021. V. 21. N 14. P. 4772. doi: 10.3390/s21144772
8. Patel P., Nandu M., Raut P. Initialization of weights in neural networks // International Journal of Scientific Development and Research. 2018. V. 3. N 11. P. 73–79.
9. Narkhede M.N., Bartakke P.P., Sutaone M.S. A review on weight initialization strategies for neural networks // Artificial Intelligence Review. 2022. V. 55. N 1. P. 291–322. doi: 10.1007/s10462-021-10033-z
10. Дайзенрот М.П., Альдо Фейзал А., Чен Сунь Он. Математика в машинном обучении. Докопайся до сути. СПб.: Питер, 2025. 507 с.
11. Шолле Ф. Глубокое обучение на Python. СПб.: Питер, 2025. 576 с.
12. Садовников П. Что мы знаем о ландшафте функции потерь в машинном обучении? // Хабр. 2018.
13. Kodryan M.S. Training Dynamics and Loss Landscape of Neural Networks with Scale-Invariant Parameters. PhD Dissertation for the purpose of obtaining academic degree Doctor of Philosophy in Computer Science. Moscow, National Research University Higher School of Economics, 2023, 103 p.
14. Толстых В.Н., Ершов В.Н., Коржевин М.А. Обход критических точек при обучении нейросети // Математические методы и модели в высокотехнологичном производстве, ГУАП. 2025. Часть 2. С. 127–129.
15. Tolstykh V.N., Yershov V.N. Local minima of error function in neural networks // Proc. of the Wave Electronics and its Application in Information and Telecommunication Systems (WECONF). 2026. P. 1–4. doi: 10.1109/weconf69969.2026.11553642

Авторы

Толстых Виктор Николаевич — кандидат технических наук, доцент, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 19000, Российская Федерация, [sc 59170433000](https://orcid.org/0009-0003-0202-5500), <https://orcid.org/0009-0003-0202-5500>, victor.n.tolstykh@gmail.com

Ершов Владимир Николаевич – кандидат физико-математических наук, доцент, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 19000, Российская Федерация, [sc 8695122100](https://orcid.org/0000-0002-9830-2159), <https://orcid.org/0000-0002-9830-2159>, vyershov@guap.ru

Статья поступила в редакцию 03.07.2026
Одобрена после рецензирования 19.08.2026
Принята к печати 25.09.2026

References

1. Pascanu R., Mikolov T., Bengio Y. Understanding the exploding gradient problem. *CoRR*, 2012, vol. 2, no. 417, pp. 1.
2. Nikolenko S., Kadurin A., Arkhangelskaya E. *Deep Learning. Diving into the World of Neural Networks*. St. Petersburg, Piter Publ., 2020, 480 p. (in Russian)
3. Li Y., Yuan Y. Convergence analysis of two-layer neural networks with ReLU activation. *Proc. of the 31st Conference on Neural Information Processing Systems*, 2017.
4. Gavin H.P. The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems. *Department of Civil and Environmental Engineering Duke University*, 2026, pp. 1–22.
5. Julius, Mahale G., Sumana T., Adityakrishna C.S. On the modeling of error functions as high dimensional landscapes for weight initialization in learning networks. *Proc. of the International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS)*, 2016, pp. 202–210. doi: 10.1109/samos.2016.7818349
6. Bociort F., Serebriakov A., Braat J. Local optimization strategies to escape from poor local minima. *International Optical Design Conference*, 2002, pp. ITuA5. doi: 10.1364/iode.2002.itua5
7. Rudd-Orthner R.N.M., Mihaylova L. Deep ConvNet: Non-random weight initialization for repeatable determinism, examined with FSGM. *Sensors*, 2021, vol. 21, no. 14, pp. 4772. doi: 10.3390/s21144772
8. Patel P., Nandu M., Raut P. Initialization of weights in neural networks. *International Journal of Scientific Development and Research*, 2018, vol. 3, no. 11, pp. 73–79.
9. Narkhede M.N., Bartakke P.P., Sutaone M.S. A review on weight initialization strategies for neural networks. *Artificial Intelligence Review*, 2022, vol. 55, no. 1, pp. 291–322. doi: 10.1007/s10462-021-10033-z
10. Deisenroth M.P., Aldo Faisal A., Cheng Soon Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020, 371 p.
11. Chollet F. *Deep Learning with Python*. Simon and Schuster, 2021, 478 p.
12. Sadovnikov P. What do we know about the loss function landscape in machine learning? *Habr*, 2018. (in Russian)
13. Kodryan M.S. *Training Dynamics and Loss Landscape of Neural Networks with Scale-Invariant Parameters*. PhD Dissertation for the purpose of obtaining academic degree Doctor of Philosophy in Computer Science. Moscow, National Research University Higher School of Economics, 2023, 103 p.
14. Tolstykh V.N., Ershov V.N., Korzhev M.A. Bypassing critical points during the neural network training. *Mathematical Methods and Models in High-Tech Manufacturing. Collection of Abstracts from the 5th International Forum*, 2025, part 2, pp. 127–129. (in Russian)
15. Tolstykh V.N., Yershov V.N. Local minima of error function in neural networks. *Proc. of the Wave Electronics and its Application in Information and Telecommunication Systems (WECONF)*, 2026, pp. 1–4. doi: 10.1109/weconf69969.2026.11553642

Aurhors

Victor N. Tolstykh — PhD, Associate Professor, Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, 190000, Russian Federation, [sc 59170433000](https://orcid.org/0009-0003-0202-5500), <https://orcid.org/0009-0003-0202-5500>, victor.n.tolstykh@gmail.com

Vladimir N. Yershov – PhD (Physics & Mathematics), Associate Professor, Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, 190000, Russian Federation, [sc 8695122100](https://orcid.org/0000-0002-9830-2159), <https://orcid.org/0000-0002-9830-2159>, vyershov@guap.ru

Received 03.07.2026
Approved after reviewing 19.08.2026
Accepted 25.09.2026



Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»