

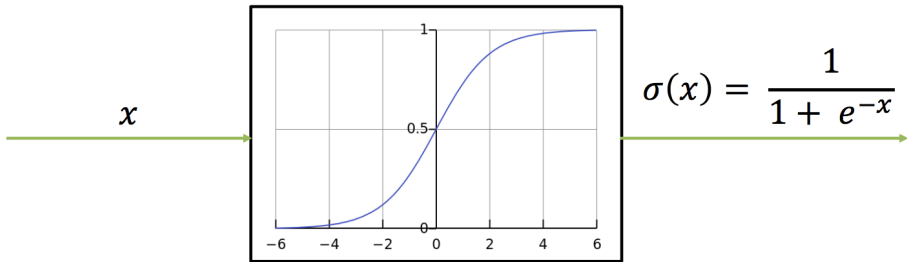
Глубокое обучение

Коломейцева Катерина

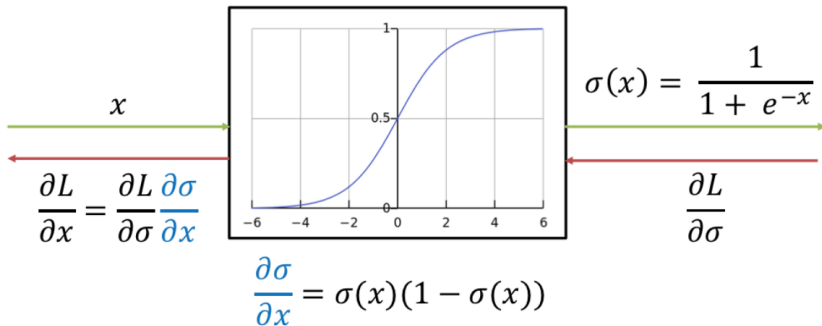
Занятие 3: эвристики для обучения нейросеток

Какими бывают функции активации
и как через них пробросить
градиент

Sigmoid activation



Sigmoid activation



Паралич сети

- В случае сигмоиды $\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$
- Сигмоида принимает значения на отрезке $[0; 1]$, значит максимальное значение её производной это $\frac{1}{4}$
- Если сеть очень глубокая, происходит **затухание градиента**
- Градиент затухает экспоненциально \Rightarrow сходимость замедляется, более ранние веса обновляются дольше, более глубокие веса быстрее \Rightarrow значение градиента становится ещё меньше \Rightarrow наступает **паралич сети**
- В сетях с небольшим числом слоёв этот эффект незаметен

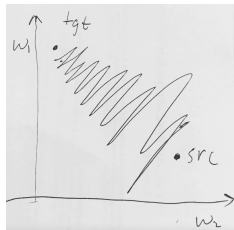
Центрирование

- Сигмоида не центрирована относительно нуля
- Выход слоя мы обычно находим как $o_i = \sigma(h_i)$, он всегда положительный, значит вектор градиента по весам, идущим на вход в текущий нейрон, тоже положительный \Rightarrow они веса обновляются в одинаковом направлении (либо все уменьшаются, либо все увеличиваются)
- Сходимость идёт медленнее и зигзагообразно, но идёт

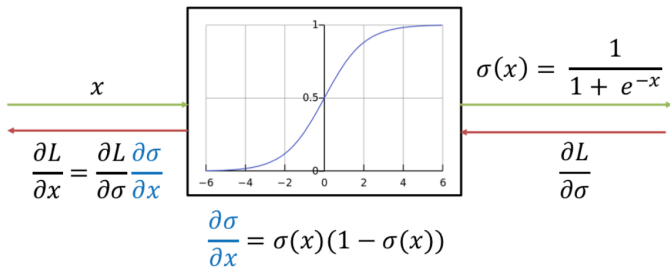
Центрирование

$$f = \sum w_i x_i + b$$
$$\frac{df}{dw_i} = x_i$$
$$\frac{dL}{dw_i} = \frac{dL}{df} \frac{df}{dw_i} = \frac{dL}{df} x_i$$

because $x_i > 0$, the gradient $\frac{dL}{dw_i}$ always has the same sign as $\frac{dL}{df}$ (all positive or all negative). For every w_i the sign is the same

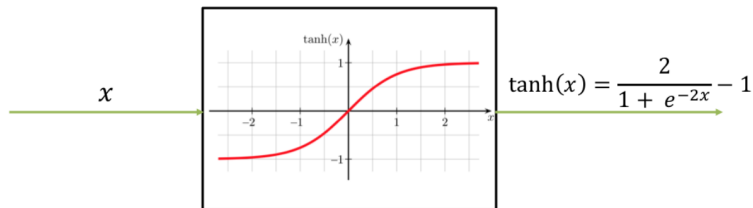


Sigmoid activation



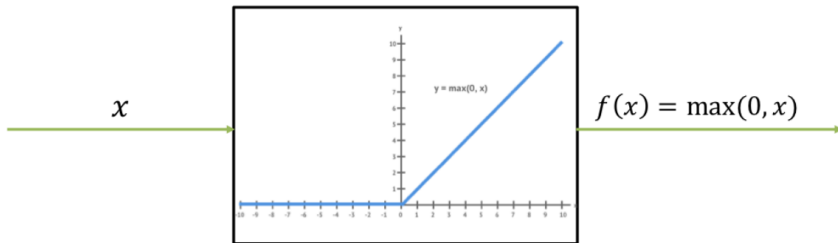
- Способствует затуханию градиента
- Не центрирована относительно нуля
- Вычислять e^x дорого

Tanh activation



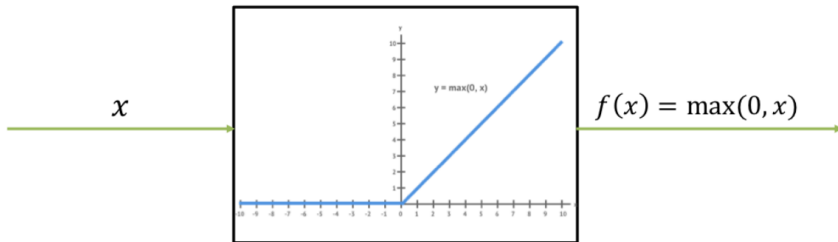
- Центрирован относительно нуля
- Всё ещё похож на сигмоиду
- $f'(x) = 1 - f(x)^2 \Rightarrow$ затухание градиента

ReLU activation



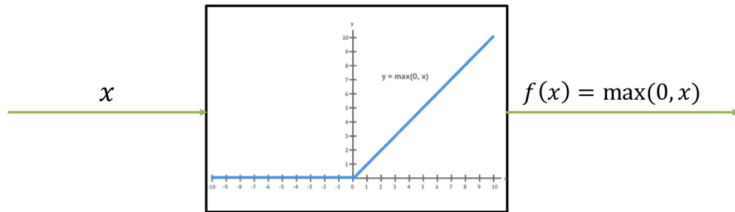
- Быстро вычисляется
- Градиент не затухает
- Сходимость сеток ускоряется

ReLU activation



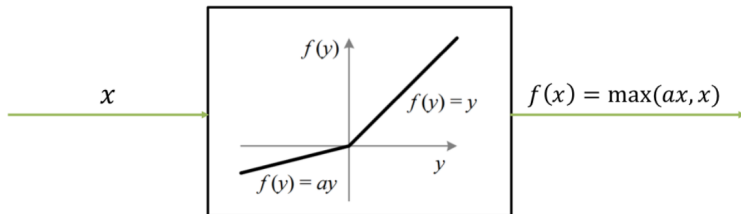
- Сетка может умереть, если активация занулитися на всех нейронах
- Не центрирован относительно нуля

Зануление ReLU



- $f(x) = \max(0, w_0 + w_1 \cdot h_1 + \dots + w_k \cdot h_k)$
- Если w_0 инициализировано большим отрицательным числом, нейрон сразу умирает \Rightarrow надо аккуратно инициализировать веса

Leaky ReLU activation



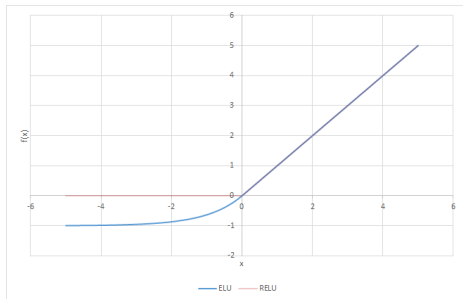
- Как ReLU, но не умирает, всё ещё легко считается
- Производная может быть любого знака
- Важно, чтобы $a \neq 1$, иначе линейность

Что же выбрать

- Обычно начинают с $ReLU$, если сетка умирает, берут $LeakyReLU$
- $ReLU$ — стандартный выбор для свёрточных сетей
- В рекуррентных сетках чаще всего предпочитается \tanh
- На самом деле это не очень важно, нужно держать в голове свойства функций, о которых выше шла речь и понимать, что от перебора функций обычно выигрыш в качестве довольно низкий
- Но есть и исключения ...

Краткий обзор функций активаций: <https://arxiv.org/pdf/1804.02763.pdf>

ELU activation



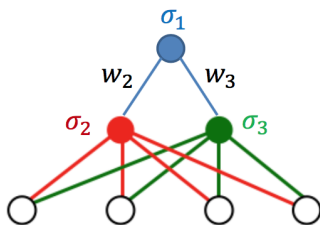
- ELU улучшает сходимость для глубоких сетей

$$f(x) = \begin{cases} x, & x \geq 0 \\ \alpha \cdot (e^x - 1), & x < 0 \end{cases}$$

<https://arxiv.org/pdf/1511.07289.pdf>

Инициализация весов

Инициализация весов

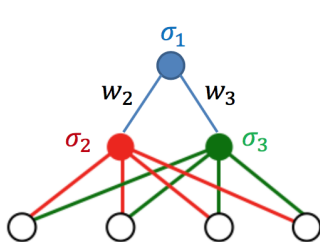


$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \sigma_1} \sigma_1 (1 - \sigma_1) \sigma_2$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial \sigma_1} \sigma_1 (1 - \sigma_1) \sigma_3$$

- Что будет, если инициализировать веса нулями?

Инициализация весов

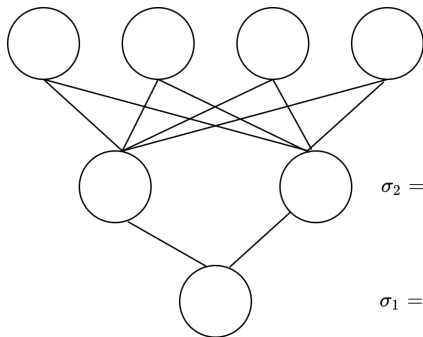


$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \sigma_1} \sigma_1 (1 - \sigma_1) \sigma_2$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial \sigma_1} \sigma_1 (1 - \sigma_1) \sigma_3$$

- Что будет, если инициализировать веса нулями?
- W_2 и W_3 будут обновляться одинаково

Инициализация весов



$$\sigma_2 = \sigma(W_1 X + b_1)$$

$$\sigma_1 = f(W_2 \sigma_2 + b_2)$$

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial \sigma_1} \sigma_1 (1 - \sigma_1) \sigma_2$$

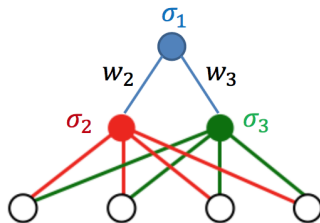
$$W_2 = W_2 - \alpha \frac{\partial L}{\partial W_2}$$

- If you have sigmoid activation $g(0) \neq 0$ $g(0) \neq 0$ then it will cause weights to move "together", limiting the power of back-propagation to search the entire space

- If you have tanhtanh or ReLu activation $g(0) = 0$ $g(0)=0$ then all the outputs will be 0, and the gradients for the weights will always be 0.

Если веса инициализируются нулями, то выходы слоев σ_1 , σ_2 либо нулевые, либо константные, что ведет к такому же обновлению весов

Инициализация весов



$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \sigma_1} \sigma_1 (1 - \sigma_1) \sigma_2$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial \sigma_1} \sigma_1 (1 - \sigma_1) \sigma_3$$

- Хочется уничтожить симметрию
- Обычно инициализируют маленькими случайными числами из какого-то распределения (нормальное, равномерное)

Инициализация весов

- Наши признаки X пришли к нам из какого-то распределения
- Выход слоя $f(XW)$ будет принадлежать другому распределению
- Если инициализировать веса неправильно, дисперсия распределения може от слоя к слою затухать (сигнал будет теряться) либо наоборот, возрастать (сигнал будет рассеиваться)
- Эмпирически было выяснено, что это может портить сходимость для глубоких сетей
- Хочется контролировать дисперсию

Инициализация весов

- Посмотрим на выход нейрона перед активацией:

$$h_i = w_0 + \sum_{i=1}^{n_{in}} w_i x_i$$

- Дисперсия h_i выражается через дисперсии x и w
- Она не зависит от константы w_0
- Будем считать, что веса $w_1, \dots, w_k \sim iid$, наблюдения $x_1, \dots, x_n \sim iid$, а ещё x_i и w_i независимы между собой

Инициализация весов (симметричный случай)

$$\begin{aligned}\text{Var}(h_i) &= \text{Var}\left(\sum_{i=1}^{n_{in}} w_i x_i\right) = \sum_{i=1}^{n_{in}} \text{Var}(w_i x_i) = \\ &= \sum_{i=1}^{n_{in}} \left([\text{E}(x_i)]^2 \cdot \text{Var}(w_i) + [\text{E}(w_i)]^2 \cdot \text{Var}(x_i) + \text{Var}(x_i) \cdot \text{Var}(w_i) \right) =\end{aligned}$$

Инициализация весов (симметричный случай)

$$\begin{aligned}\text{Var}(h_i) &= \text{Var}\left(\sum_{i=1}^{n_{in}} w_i x_i\right) = \sum_{i=1}^{n_{in}} \text{Var}(w_i x_i) = \\ &= \sum_{i=1}^{n_{in}} \left([\mathbf{E}(x_i)]^2 \cdot \text{Var}(w_i) + [\mathbf{E}(w_i)]^2 \cdot \text{Var}(x_i) + \text{Var}(x_i) \cdot \text{Var}(w_i) \right) =\end{aligned}$$

- Если функция активации симметричная ($f(z) + f(-z) = 1$), тогда $E(x_i) = 0$.
- Будем инициализировать веса с нулевым средним, тогда $E(w_i) = 0$.

Инициализация весов (симметричный случай)

$$\begin{aligned}\text{Var}(h_i) &= \text{Var}\left(\sum_{i=1}^{n_{in}} w_i x_i\right) = \sum_{i=1}^{n_{in}} \text{Var}(w_i x_i) = \\ &= \sum_{i=1}^{n_{in}} \left([\mathbf{E}(x_i)]^2 \cdot \text{Var}(w_i) + [\mathbf{E}(w_i)]^2 \cdot \text{Var}(x_i) + \text{Var}(x_i) \cdot \text{Var}(w_i) \right) = \\ &= \sum_{i=1}^{n_{in}} \text{Var}(x_i) \cdot \text{Var}(w_i)\end{aligned}$$

- Если функция активации симметричная ($f(z) + f(-z) = 1$), тогда $\mathbf{E}(x_i) = 0$.
- Будем инициализировать веса с нулевым средним, тогда $\mathbf{E}(w_i) = 0$.

Инициализация весов (симметричный случай)

$$\begin{aligned}\text{Var}(h_i) &= \text{Var}\left(\sum_{i=1}^{n_{in}} w_i x_i\right) = \sum_{i=1}^{n_{in}} \text{Var}(w_i x_i) = \\ &= \sum_{i=1}^{n_{in}} \left([\text{E}(x_i)]^2 \cdot \text{Var}(w_i) + [\text{E}(w_i)]^2 \cdot \text{Var}(x_i) + \text{Var}(x_i) \cdot \text{Var}(w_i) \right) = \\ &= \sum_{i=1}^{n_{in}} \text{Var}(x_i) \cdot \text{Var}(w_i) = \text{Var}(x) \cdot [n_{in} \cdot \text{Var}(w)]\end{aligned}$$

Инициализация весов (симметричный случай)

$$\begin{aligned}\text{Var}(h_i) &= \text{Var}\left(\sum_{i=1}^{n_{in}} w_i x_i\right) = \sum_{i=1}^{n_{in}} \text{Var}(w_i x_i) = \\ &= \sum_{i=1}^{n_{in}} \left([\text{E}(x_i)]^2 \cdot \text{Var}(w_i) + [\text{E}(w_i)]^2 \cdot \text{Var}(x_i) + \text{Var}(x_i) \cdot \text{Var}(w_i) \right) = \\ &= \sum_{i=1}^{n_{in}} \text{Var}(x_i) \cdot \text{Var}(w_i) = \text{Var}(x) \cdot \underbrace{[n_{in} \cdot \text{Var}(w)]}_{=1}\end{aligned}$$

Плохая инициализация весов

Пусть

$$w_i \sim U \left[-\frac{1}{\sqrt{n_{in}}}; \frac{1}{\sqrt{n_{in}}} \right],$$

тогда

$$\text{Var}(w_i) = \frac{1}{12} \cdot \left(\frac{1}{\sqrt{n_{in}}} + \frac{1}{\sqrt{n_{in}}} \right)^2 = \frac{1}{3n_{in}} \Rightarrow \text{Var}(h_i) = \frac{\text{Var}(x)}{3}$$

Получаем затухание!

Немного лучше

Пусть

$$w_i \sim U \left[-\frac{\sqrt{3}}{\sqrt{n_{in}}}; \frac{\sqrt{3}}{\sqrt{n_{in}}} \right],$$

тогда

$$\text{Var}(w_i) = \frac{1}{12} \cdot \left(\frac{\sqrt{3}}{\sqrt{n_{in}}} + \frac{\sqrt{3}}{\sqrt{n_{in}}} \right)^2 = \frac{1}{n_{in}} \Rightarrow \text{Var}(h_i) = \text{Var}(x)$$

Немного лучше

Пусть

$$w_i \sim U \left[-\frac{\sqrt{3}}{\sqrt{n_{in}}}; \frac{\sqrt{3}}{\sqrt{n_{in}}} \right],$$

тогда

$$\text{Var}(w_i) = \frac{1}{12} \cdot \left(\frac{\sqrt{3}}{\sqrt{n_{in}}} + \frac{\sqrt{3}}{\sqrt{n_{in}}} \right)^2 = \frac{1}{n_{in}} \Rightarrow \text{Var}(h_i) = \text{Var}(x)$$

При forward pass на вход идёт n_{in} наблюдений, при backward pass на вход идёт n_{out} градиентов \Rightarrow **канал с дисперсией может быть непостоянным, если число весов от слоя к слою сильно колеблется**

Инициализация Xavier Glorot

Для неодинаковых размеров слоёв невозможно удовлетворить обоим условиям, поэтому обычно усредняют:

$$w_i \sim U \left[-\frac{\sqrt{6}}{\sqrt{n_{out} + n_{in}}}; \frac{\sqrt{6}}{\sqrt{n_{out} + n_{in}}} \right],$$

Такая инициализация называется **инициализацией Xavier Glorot**

<http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf>

Инициализация Хе Kaiming

$$\begin{aligned}\text{Var}(h_i) &= \text{Var}\left(\sum_{i=1}^{n_{in}} w_i x_i\right) = \sum_{i=1}^{n_{in}} \text{Var}(w_i x_i) = \\ &= \sum_{i=1}^{n_{in}} [\text{E}(x_i)]^2 \cdot \text{Var}(w_i) + [\text{E}(w_i)]^2 \cdot \text{Var}(x_i) + \text{Var}(x_i) \cdot \text{Var}(w_i)]\end{aligned}$$

- Когда нет симметрии, можно занулить только второе слагаемое

Инициализация Хе Kaiming

$$\begin{aligned}\text{Var}(h_i) &= \text{Var}\left(\sum_{i=1}^{n_{in}} w_i x_i\right) = \sum_{i=1}^{n_{in}} \text{Var}(w_i x_i) = \\ &= \sum_{i=1}^{n_{in}} [\text{E}(x_i)]^2 \cdot \text{Var}(w_i) + [\text{E}(w_i)]^2 \cdot \text{Var}(x_i) + \text{Var}(x_i) \cdot \text{Var}(w_i) = \\ &= \sum_{i=1}^{n_{in}} [\text{E}(x_i)]^2 \cdot \text{Var}(w_i) + \text{Var}(x_i) \cdot \text{Var}(w_i) = \sum_{i=1}^{n_{in}} \text{Var}(w_i) \cdot E(x_i^2)\end{aligned}$$

- Когда нет симметрии, можно занулить только второе слагаемое

Инициализация Хе Kaiming

$$\begin{aligned}\text{Var}(h_i) &= \text{Var}\left(\sum_{i=1}^{n_{in}} w_i x_i\right) = \sum_{i=1}^{n_{in}} \text{Var}(w_i x_i) = \\&= \sum_{i=1}^{n_{in}} [\text{E}(x_i)]^2 \cdot \text{Var}(w_i) + [\text{E}(w_i)]^2 \cdot \text{Var}(x_i) + \text{Var}(x_i) \cdot \text{Var}(w_i) = \\&= \sum_{i=1}^{n_{in}} [\text{E}(x_i)]^2 \cdot \text{Var}(w_i) + \text{Var}(x_i) \cdot \text{Var}(w_i) = \sum_{i=1}^{n_{in}} \text{Var}(w_i) \cdot E(x_i^2) = \\&= E(x^2) \cdot [n_{in} \cdot \text{Var}(w)]\end{aligned}$$

Инициализация Хе Kaiming

$$\begin{aligned}\mathbf{Var}(h_i) &= E(x_i^2) \cdot [n_{in} \cdot \mathbf{Var}(w)] \\ x_i &= \max(0; h_{i-1})\end{aligned}$$

Инициализация Хе Kaiming

$$\begin{aligned}\text{Var}(h_i) &= E(x_i^2) \cdot [n_{in} \cdot \text{Var}(w)] \\ x_i &= \max(0; h_{i-1})\end{aligned}$$

Пусть w_{i-1} симметрично распределён относительно нуля и $b_{i-1} = 0$. Тогда h_{i-1} также симметрично распределён относительно нуля в силу того, что $h_{i-1} = w_{i-1} \cdot \text{ReLU}(h_{i-2})$. А если h_{i-1} также симметрично распределён относительно нуля, то

$$E(x_i^2) = \frac{1}{2} \cdot \text{Var}(h_{i-1})$$

Инициализация Хе Kaiming

Если

$$E(x_i^2) = \frac{1}{2} \cdot \text{Var}(h_{i-1})$$

Тогда

$$\text{Var}(h_i) = \frac{1}{2} \cdot \text{Var}(h_{i-1}) \cdot [n_{in} \cdot \text{Var}(w)]$$

$$\text{Var}(w_i) = \frac{2}{n_{in}}$$

Итого: инициализируем веса нормальным распределением с нулевым мат. ожиданием и дисперсией $\frac{2}{n_{in}}$:

$$N(0, \frac{2}{n_{in}})$$

Инициализация Хе - пояснения к формулам выше

h_{i-1} has zero mean and has a symmetrical distribution around zero.

$$\begin{aligned}\text{Var}(h_{i-1}) &= \mathbb{E}[h_{i-1}^2] \\ &= \mathbb{E}[h_{i-1}^2 \mid h_{i-1} > 0] \Pr[h_{i-1} > 0] + \mathbb{E}[h_{i-1}^2 \mid h_{i-1} < 0] \Pr[h_{i-1} < 0] \\ &= 2\mathbb{E}[h_{i-1}^2 \mid h_{i-1} > 0] \Pr[h_{i-1} > 0]\end{aligned}$$

Since

$$x_i = \max(0, h_{i-1})$$

Hence

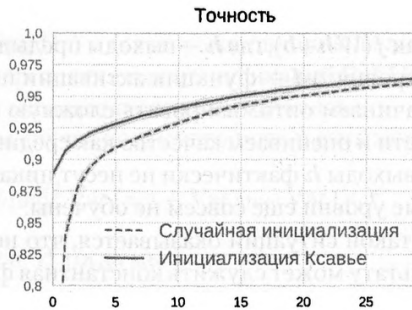
$$x_i^2 = \begin{cases} h_{i-1}^2 & , h_{i-1} > 0 \\ 0 & , h_{i-1} \leq 0 \end{cases}$$

$$\implies \mathbb{E}[x_i^2] = \mathbb{E}[h_{i-1}^2 \mid h_{i-1} > 0] \Pr[h_{i-1} > 0] = \frac{1}{2} \text{Var}(h_{i-1})$$

Краткие итоги

- Для симметричных функций с нулевым средним используйте инициализацию Ксавье `init="glorot_uniform"` или
- Для ReLU и им подобным инициализацию Хе `init="he_uniform"` или `init="he_nomal"`
- Эти две инициализации корректируют параметры распределений в зависимости от входа и выхода слоя так, чтобы поддерживать дисперсию равной единице

Эксперимент с MNIST



Источник: Николенко, страница 149