JOHANNES KEPLER
UNIVERSITY LINZ

JOHANNES KEPLER
UNIVERSITY LINZ

Deep Learning and Neural Networks I: 9. Initialization

Günter Klambauer
LIT AI Lab & Institute for Machine Learning

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Strasse 69
4040 Linz, Austria
jku.at

Initializing neural networks

- Neural networks are trained iteratively starting with initial weights

$$\boldsymbol{w}^{(0)} \to \ldots \to \boldsymbol{w}^{(t)}$$

- Choice of initial weights matters for learning (strong effect!)
  - Initial point determines whether algorithm converges at all

Recap on prob. theory.

Example of discrete distribution: Part 1

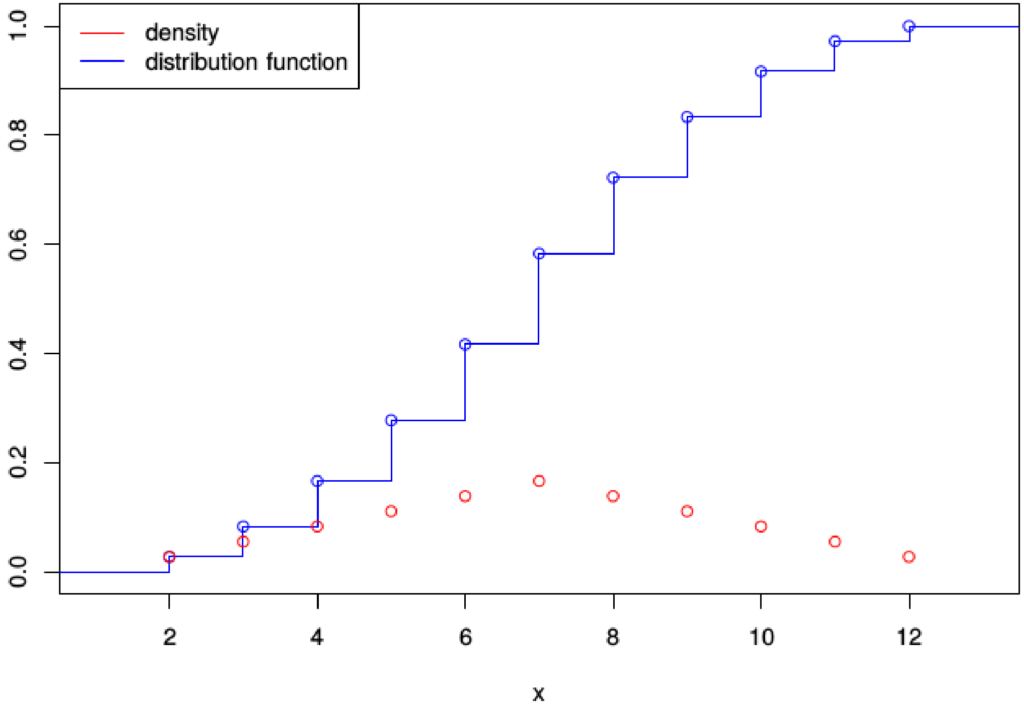
Suppose we roll two dice. Then we have $\Omega = \{(i,j) \mid i,j = 1,\ldots,6\}$, i.e. 36 different outcomes. All 36 elementary events have the same probability $\frac{1}{36}$. If Z is the function that maps each outcome to the sum of the two numbers, we see, for example,

$$Z^{-1}(\{6\}) = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}.$$

Then the distribution density of Z is given as

x	2	3	4	5	6	7	8	9	10	11	12
$p_Z(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Example of discrete distribution: Part 2



Example of continuous distribution: Part 1

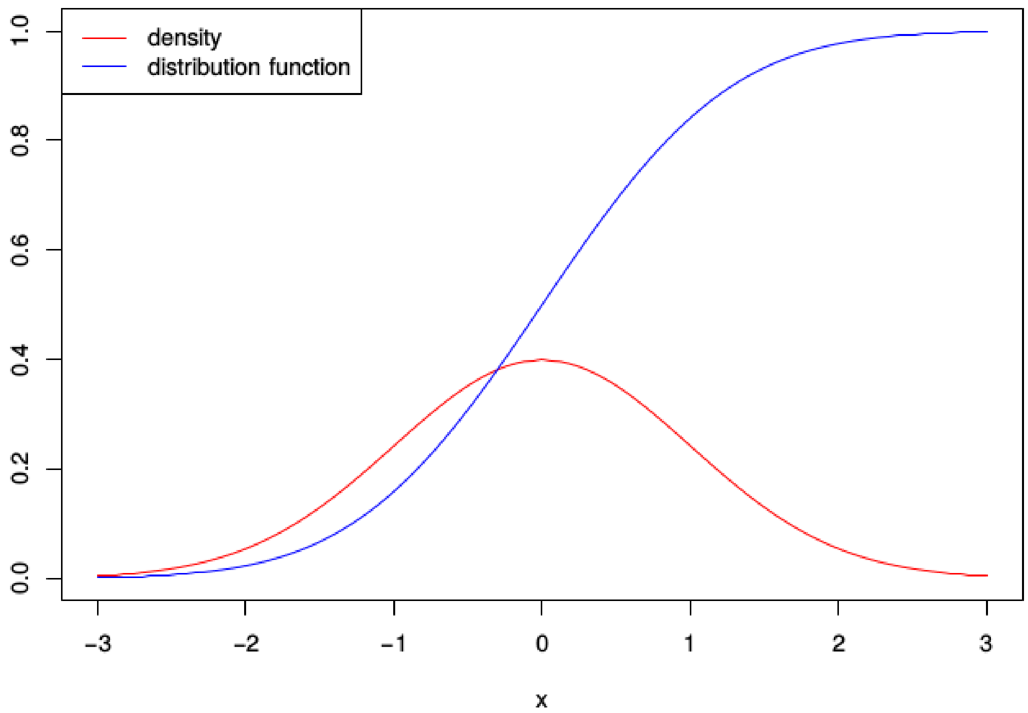
Every industrial process has some (minor) random components. Even if the deviations of the produced pieces are very small, we can speak of a “random experiment”.

Suppose we consider a certain physical parameter of the produced pieces (lengths, weights, etc.) as the random variable of interest. Such random variables are often normally (Gaussian) distributed, i.e.

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for some μ and σ.

Example of continuous distribution: Part 2



Expected value of a random variable

The expected value, aka expectation (value), of a random variable is its average outcome (in a large number of trials). It is not necessarily the most probable value!

- Discrete distribution:

$$\mathsf{E}(X) = \sum_{x} x \cdot p_X(x)$$

- Continuous distribution:

$$\mathsf{E}(X) = \int_{x} x \cdot p_X(x) dx$$

It is quite common to denote the expected value by μ_X (or simply μ if it is clear which random variable/distribution is meant).

Higher moments: Part 1

- For a random variable X, the k-th moment is defined as follows:
$$m_X^k = \mathsf{E}(X^k) = \int x^k \cdot p_X(x)dx$$
- The k-th central moment is defined as follows:
$$\mu_X^k = E\big((X - \mathsf{E}(X))^k\big) = \int (x - \mathsf{E}(X))^k \cdot p_X(x)dx$$
- The second central moment,
$$E\big((X - \mathsf{E}(X))^2\big) = \int (x - \mathsf{E}(X))^2 \cdot p_X(x)dx$$
is called variance of X and denoted with $\mathsf{Var}(X)$ or σ_Z^2.
- As in UNIT 1: $\sqrt{\mathsf{Var}(X)}$ is called the standard deviation of X.

Some fundamental rules:

Let X, Y, Z be random variables.

- $\mathsf{E}(\alpha \cdot Z + \beta) = \alpha \cdot \mathsf{E}(Z) + \beta$
- $\mathsf{E}(\alpha \cdot X + \beta \cdot Y) = \alpha \cdot \mathsf{E}(X) + \beta \cdot \mathsf{E}(Y)$
- If $X \leq Y$ then $\mathsf{E}(X) \leq \mathsf{E}(Y)$.
- $\mathsf{Var}(Z) = \mathsf{E}(Z^2) - \mathsf{E}(Z)^2$
- $\mathsf{Var}(\alpha \cdot Z + \beta) = \alpha^2 \cdot \mathsf{Var}(Z)$

Continuous distributions: Normal distribution: Part 1

- Suppose we have a random variable that is the sum of many independent random variables and whose expected value is μ and whose variance is σ^2. Then this random variable is distributed according to the normal distribution $\mathcal{N}(\mu, \sigma^2)$.
- Density: for $x \in \mathbb{R}$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Moments:
  - $\mathrm{E}(X) = \mu$
  - $\mathrm{Var}(X) = \sigma^2$

Central limit theorem

Suppose we have random variables $X_1, X_2, \ldots$ which are independent and identically distributed (i.i.d.); suppose they have expected value μ and variance σ^2. Let

$$X'_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

be the standardized n-th partial sum of random variables. Then the distribution of X'_n converges to $\mathcal{N}(0,1)$ as n goes to infinity. We won't formalize this further here!

Fundamental rules

Let X, Y be random variables.

- $\mathsf{Cov}(X,Y) = \mathsf{E}(X \cdot Y) - \mathsf{E}(X) \cdot \mathsf{E}(Y)$
- $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + 2 \cdot \mathsf{Cov}(X,Y) + \mathsf{Var}(Y)$
- $\mathsf{R}(X,Y) \in [-1,1]$
- $\mathsf{R}(X,Y) \in \{-1,1\}$ if and only if X and Y are linearly correlated, i.e. there exist $\alpha, \beta \in \mathbb{R}$ such that $Y = \alpha \cdot X + \beta$.

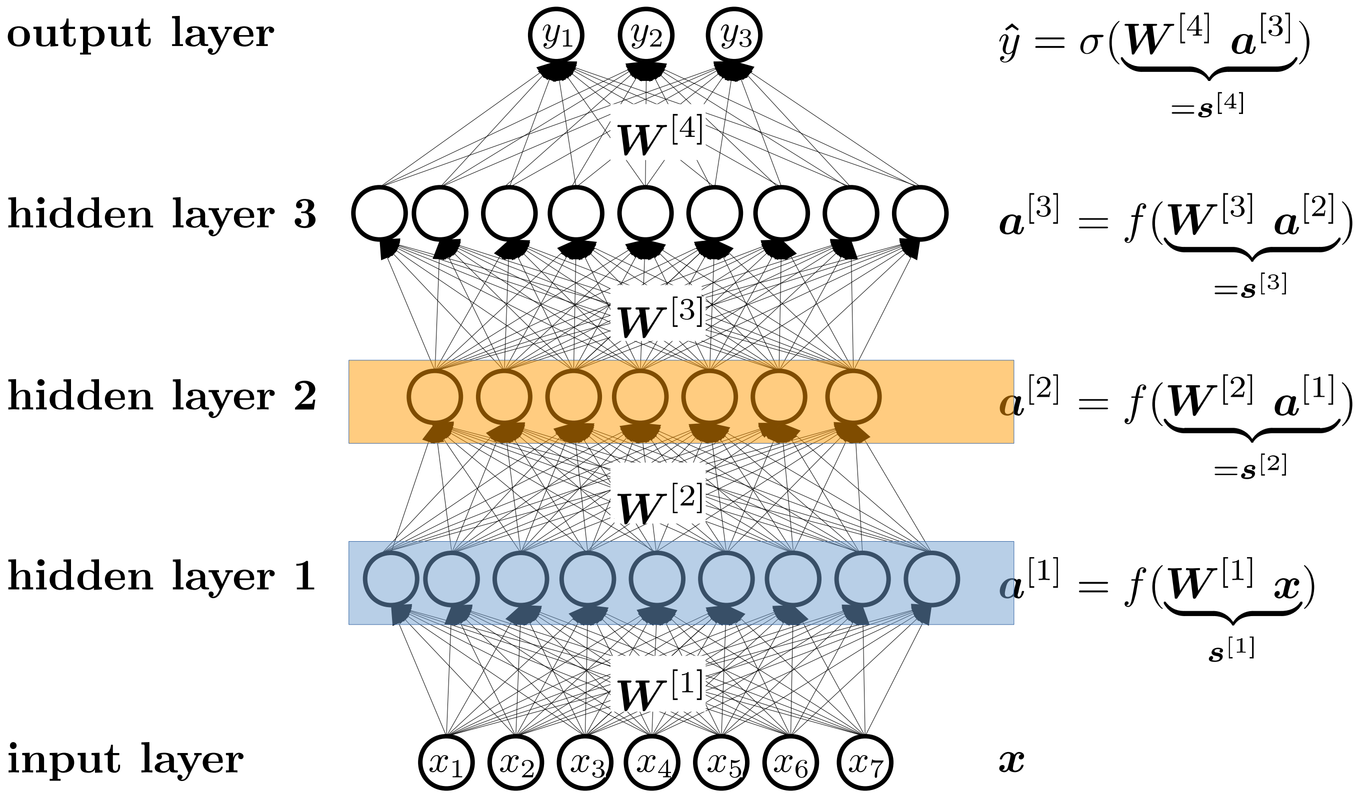
If X are Y are independent, the following holds:

- $\mathsf{E}(X \cdot Y) = \mathsf{E}(X) \cdot \mathsf{E}(Y)$
- $\mathsf{Cov}(X,Y) = 0$
- $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$

Overview

- 9.1 Breaking symmetry
  - 9.1.1 Constant initialization
  - 9.1.2 Random initialization
  - 9.1.3 Bias Weights
- 9.2 Mean field theory for initialization
  - 9.2.1 Variance Propagation
  - 9.2.2 Error Propagation
- 9.3. Non-linearities
  - 9.3.1 Propagation function
  - 9.3.2 Gain factor

Notation: focus on mapping from one layer to the next



- Number of neurons in previous **layer**: J
- Number of neurons in next **layer**: I

9.1 Breaking symmetry

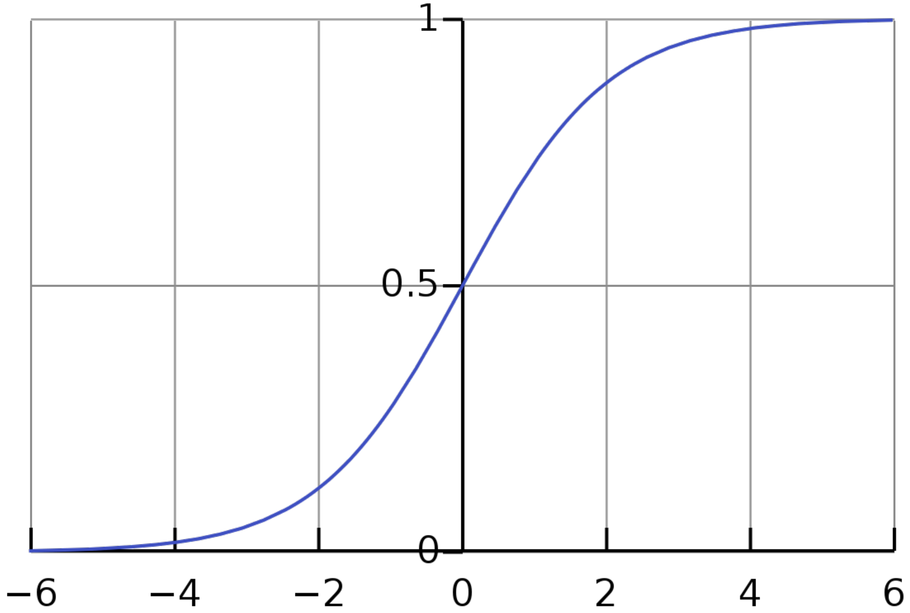
- Assume two units with same weights (and same activation functions)
  - Optimization will update weights identically
  - Same features represented in network
  - Equivalence of neurons is called “symmetry”
- Each unit initialized differently
  - Starts with random pattern in input data

9.1.2 Random initialization

- Each neuron should start with capturing distinct patterns from input
- Therefore, random initialization
  - Which distribution?
  - Which parameters (mean, var) of distribution?

9.1.2 Random initialization

- First idea: starting close to a linear network
- Sigmoid is linear close to zero
- Hence: initialize weights with small values



Uniform distribution with small interval:

$$\mathrm{W}_{ij} \sim \mathcal{U}(-\epsilon, \epsilon)$$

Normal distribution with small variance:

$$W_{ij} \sim \mathcal{N}(0, \sigma^2)$$

9.1.2 Random initialization

- Uniform distribution

$$\mathrm{W}_{ij} \sim \mathcal{U}(-\epsilon, \epsilon)$$

- Gaussian distribution

$$W_{ij} \sim \mathcal{N}(0, \sigma^2)$$

- Truncated Gaussian distribution

$$W_{ij} \sim \mathcal{N}_{\mathrm{trunc}}(0, \sigma^2)$$

Note that the variance of truncated Gaussians is lower than that of the original Gaussian!

9.2 Mean Field Theory

- Aim: investigate how a signal propagates through the network.
- Assumption: inputs can be modelled by i.i.d. random variables.
  - In practice, inputs will not be i.i.d., e.g. pixels in an image are generally highly correlated with neighboring pixels.
  - Could hold better for fully-connected networks
- Still useful to study the averaged effect of individual components in the often high dimensional inputs.
  - This can be considered as an application of *mean field theory* (effect of all components is approximated by single averaged effect)

9.2 Mean field theory: Averaged effects of neurons; Central limit theorem (CLT)

$x_1 \sim V_1(\mu_x, \sigma_x^2)$

$x_2 \sim V_2(\mu_x, \sigma_x^2)$

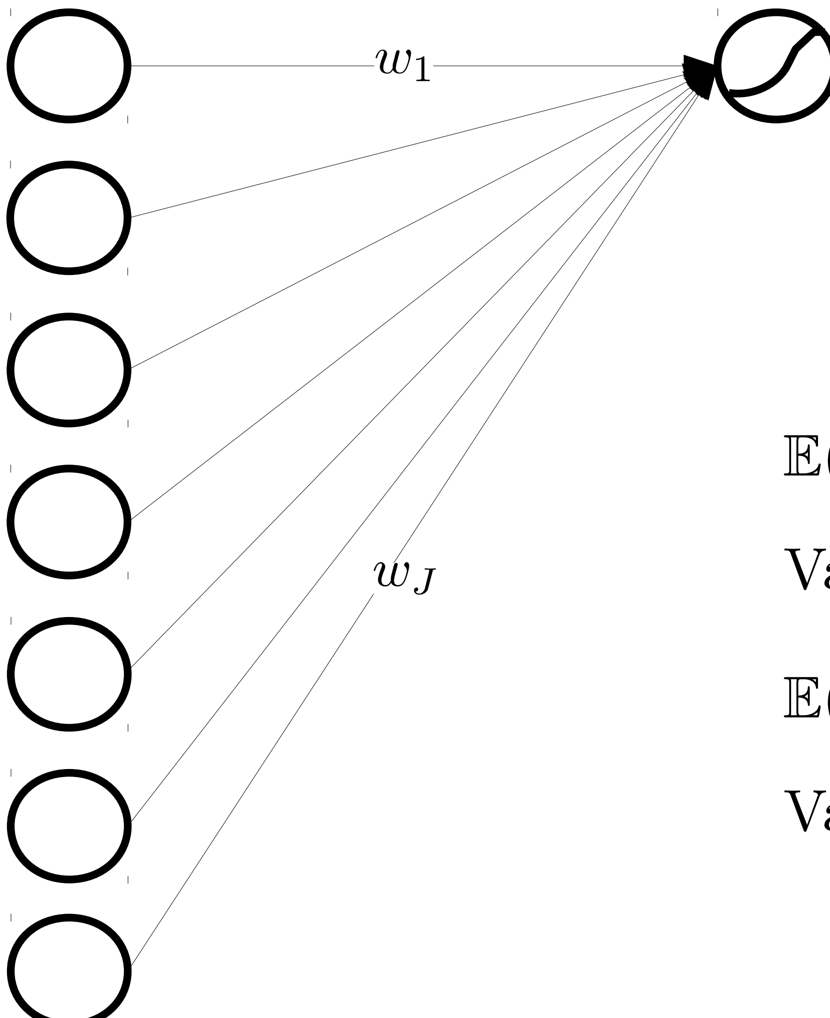
$x_3 \sim V_3(\mu_x, \sigma_x^2)$

$x_4 \sim V_4(\mu_x, \sigma_x^2)$

$x_5 \sim V_5(\mu_x, \sigma_x^2)$

$\dots$

$x_J \sim V_J(\mu_x, \sigma_x^2)$



$\boxed{s \sim \mathcal{N}(J\mu_x\mu_w, J\sigma_x^2\sigma_w^2)}$

$\boxed{y = f(s) \sim V(\tilde{\mu}, \tilde{\nu})}$

$s = \sum_{j=1}^{J} w_i x_i$

$$\mathbb{E}(x_j) = \mu_x$$

$$\mathrm{Var}(x_j) = \sigma_x^2$$

$$\mathbb{E}(w_j) = \mu_w$$

$$\mathrm{Var}(w_j) = \sigma_w^2$$

9.2 Variance propagation

- Fully-connected network.
- Mean propagation:

$$\mathbb{E}_{W_{ij},X_j}\left[\sum_j W_{ij}X_j\right] = J\mu_w\,\mathbb{E}_{X_j}[X_j]$$

- Second-moment propagation:

$$\mathbb{E}_{W_{ij},X_j}\left[\left(\sum_j W_{ij}X_j\right)^2\right] = J\sigma_w^2\,\mathbb{E}_{X_j}\left[X_j^2\right]$$

assuming that expected value of weights is zero

9.2 Variance propagation

- Second-moment propagation:

$$\mathbb{E}_{W_{ij},X_j}\left[\left(\sum_j W_{ij}X_j\right)^2\right] = J\sigma_w^2\,\mathbb{E}_{X_j}\left[X_j^2\right]$$

  Variance can blow up ($J\sigma_w^2 > 1$) or decrease through layers ($J\sigma_w^2 < 1$).

- Solution/amelioration: set variance of weights to $\sigma_w^2 = \frac{1}{J}$.

$$\mathbb{E}_{W_{ij},X_j}\left[\left(\sum_j W_{ij}X_j\right)^2\right] = J\frac{1}{J}\,\mathbb{E}_{X_j}\left[X_j^2\right] = \mathbb{E}_{X_j}\left[X_j^2\right]$$

- Result: Linear transformation does not affect variance.

9.2 Variance propagation: LeCun's initialization (LeCun, 1998)

- Already suggested in 1998 to initialize as follows:

$$W_{ij} \sim \mathcal{N}(0, \frac{1}{J}) \qquad\qquad W_{ij} \sim \mathcal{U}(-\sqrt{\frac{3}{J}}, \sqrt{\frac{3}{J}}).$$

9.2. Error Propagation

- We now use similar thoughts for the backward pass:
  - Assume deltas as i.i.d random variables

$$\mathbb{E}_{W_{ij},D_i}\left[\sum_i D_i W_{ij}\right] = I\mu_w\,\mathbb{E}\left[D_i\right] = 0$$

$$\mathbb{E}_{W_{ij},D_i}\left[\left(\sum_i D_i W_{ij}\right)^2\right] = I\sigma_w^2\,\mathbb{E}_{D_i}\left[D_i^2\right],$$

  - Note the difference between J and I !
- Initialize weights with variance $\sigma_w^2 = \frac{1}{I}$

Note: “fan-in” and “fan-out”

- Number of incoming connections from lower layer: “fan-in” J

- Number of incoming connections from higher layer: “fan-out” I

9.2 Mean-field theory for initialization: trade-off between forward and backw.

- Glorot's initialization (Glorot and Bengio, 2010):

$$W_{ij} \sim \mathcal{N}(0, \frac{2}{J+I}) \qquad W_{ij} \sim \mathcal{U}(-\sqrt{\frac{6}{J+I}}, \sqrt{\frac{6}{J+I}})$$

9.3 Non-linearities

- So far we have considered the effect of the linear transformation
- Also activation functions change the distribution of the neuron activations
- We well consider the propagation from pre-activation in one layer to the pre-activation in the next layer

9.3 Non-linearities

- Propagation of moments with non-linearities:

$$\mathbb{E}_{W_{ij},S_j\sim\mathcal{N}(\mu_s,\sigma_s^2)}\left[\sum_j W_{ij}f(S_j)\right] = J\mu_w\,\mathbb{E}_{S_j\sim\mathcal{N}(\mu_s,\sigma_s^2)}\left[f(S_j)\right] = 0$$

$$\mathbb{E}_{W_{ij},S_j\sim\mathcal{N}(\mu_s,\sigma_s^2)}\left[\left(\sum_j W_{ij}f(S_j)\right)^2\right] = J\sigma_w^2\,\mathbb{E}_{S_j\sim\mathcal{N}(\mu_s,\sigma_s^2)}\left[f(S_j)^2\right],$$

- We again assume centered weights $\mu_w = 0$

9.3 Non-linearities

- Pre-activations are weighted sums of inputs
- Assuming wide networks, the CLT applies and the pre-activations can be considered normally-distributed: $S_i = \mathcal{N}(0, \sigma_s^2)$

$$\sigma_s^2 = J\sigma_w^2 \begin{cases} \mathbb{E}_{X_j}\left[X_j^2\right] & l = 1 \\ \mathbb{E}_{S_j \sim \mathcal{N}(0,\sigma_s^2)}\left[f(S_j)^2\right] & l > 1 \end{cases}.$$

9.3 Non-linearities

- Expressions for moments:

$$\mathbb{E}_{S_j \sim \mathcal{N}(0,\sigma_s^2)}[f(S_j)^n] = \mathbb{E}_{Z \sim \mathcal{N}(0,1)}[f(\sigma_s Z)^n] = \frac{1}{\sqrt{2\pi\sigma_s^2}} \int_{\mathbb{R}} f(x)^n e^{-\frac{x}{2\sigma_s^2}} \mathrm{d}x$$

- Propagation function:

$$F_f : \mathbb{R}^+ \to \mathbb{R}^+ : q \mapsto F_f(q) = J\sigma_w^2 \, \mathbb{E}_{Z \sim \mathcal{N}(0,1)} \left[f(\sqrt{q}Z)^2 \right]$$

9.3 Non-linearities

- Application of theory to ReLU activations:

$$\mathbb{E}[\mathrm{ReLU}(S)^2] = \mathbb{E}_{S<0}[0] + \mathbb{E}_{S\geq 0}[S^2] = \frac{1}{2}\sigma_s^2.$$

where $S \sim \mathcal{N}(0, \sigma_s^2)$

- Propagation function of ReLU:

$$F_{\mathrm{ReLU}}(q) = J\sigma_w^2 \frac{1}{2} q.$$

9.3 Non-linearities: He's initialization (2015)

- Account for the effect of ReLU:

$$J\sigma_w^2 \frac{1}{2} = 1$$

  Initialize weights that counter keep unit variance:

$$\sigma_w^2 = \frac{2}{J}.$$

- Resulting initialization:

$$W_{ij} \sim \mathcal{N}(0, \frac{2}{J}) \qquad W_{ij} \sim \mathcal{U}(-\sqrt{\frac{6}{J}}, \sqrt{\frac{6}{J}})$$

Gain factor

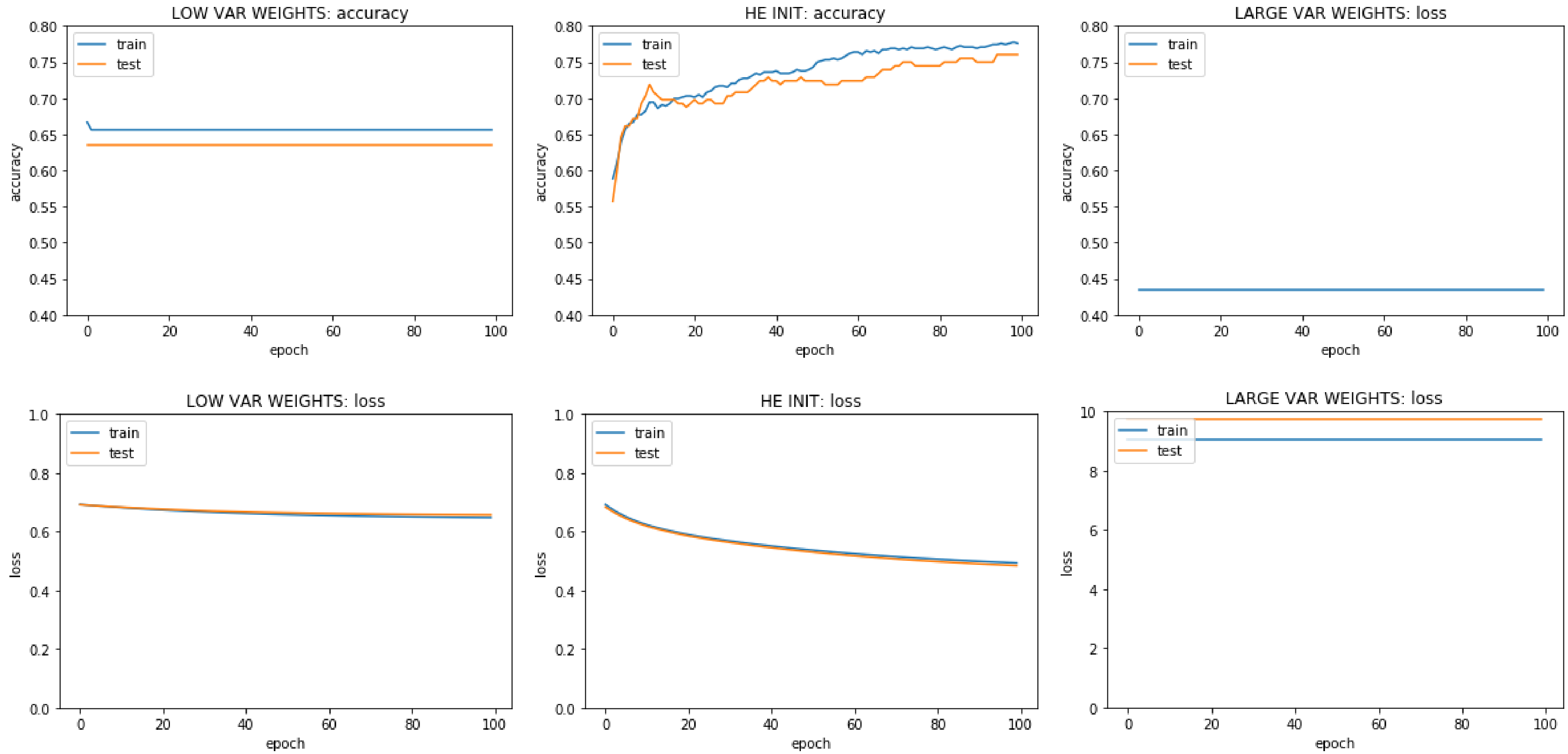
- Effect of activation function can be modelled by a positive gain factor

$$F_f(\sigma_s^2) = J\sigma_w^2\,\mathbb{E}[f(S)^2] \approx J\sigma_w^2 \frac{1}{g_f}\sigma_s^2$$

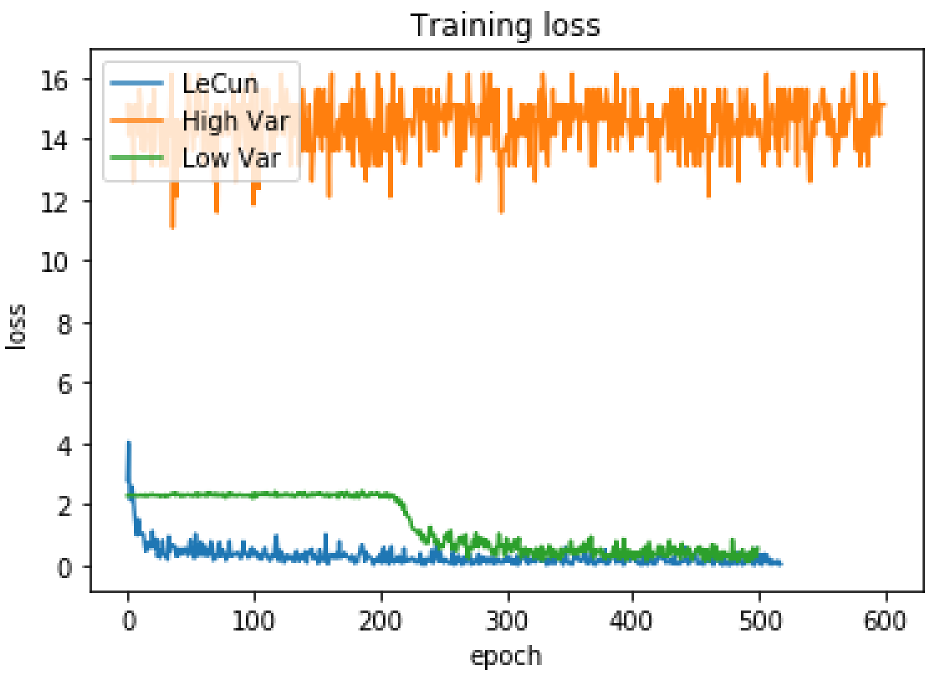
- As a result, the initial weights are sampled from a distribution with a variance

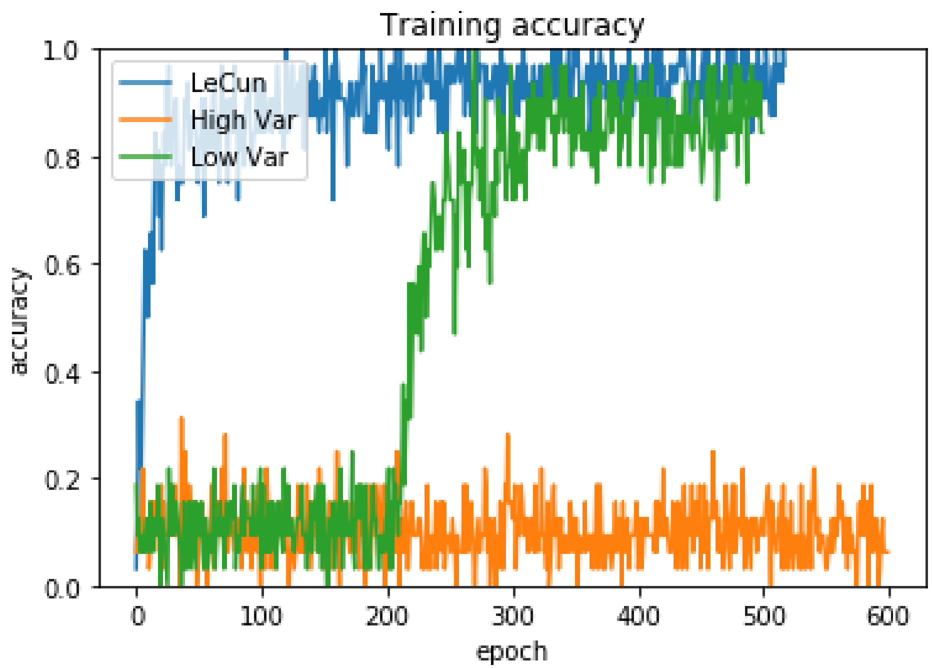
$$\sigma_w^2 = \frac{g_f}{J} \qquad \text{or} \qquad \sigma_w^2 = g_f \frac{2}{I+J}$$

Results: Fully-connected networks effect of initialization



Results: CNNS on MNIST
effect of initialization





Summary

- Overview of initialization strategies
  - Constant initialization – problems
  - Breaking symmetry
  - Variance/error propagation
  - The role of non-linearities