Jonas Glombitza
jonas.glombitza@fau.de

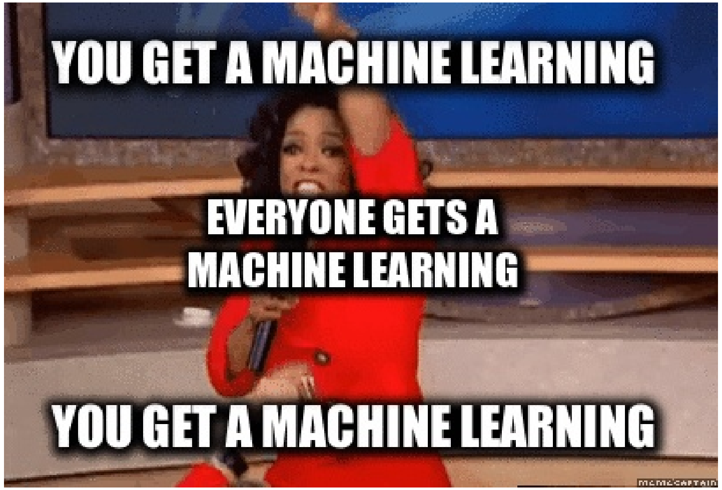
ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

FAU

Astroparticle School 2025
Waischenfeld, Fraunhofer Research Campus

https://github.com/DeepLearningForPhysicsResearchBook/deep-learning-physics

Deep Learning for Physics Research



I. Basic Methods & Techniques
II. Deep Learning Frameworks
III. Physics Examples and Applications

Time schedule for the next days

Tutorial: Introduction to deep learning — **Tuesday 1:15h**

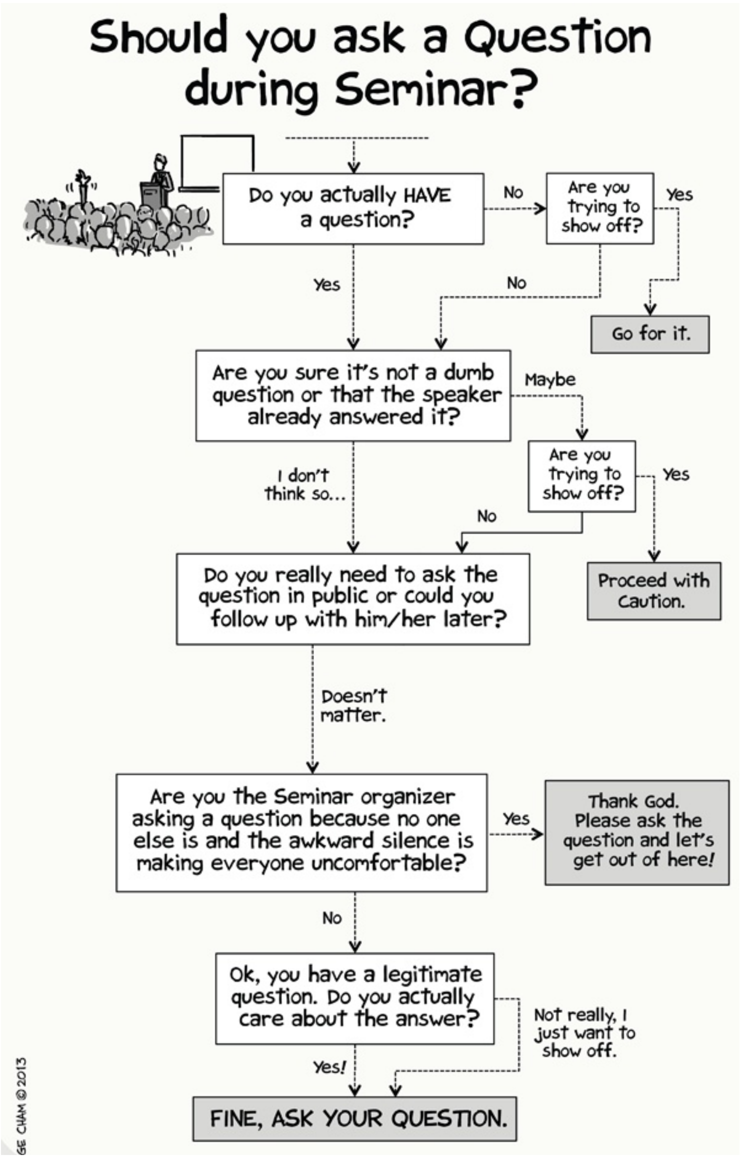
- Training of deep neural networks
- Interactive training of neural networks

Hands-on — **Friday 1:30h**

- Convolutional neural network
- machine learning frameworks: Keras
- Implementation of deep neural networks

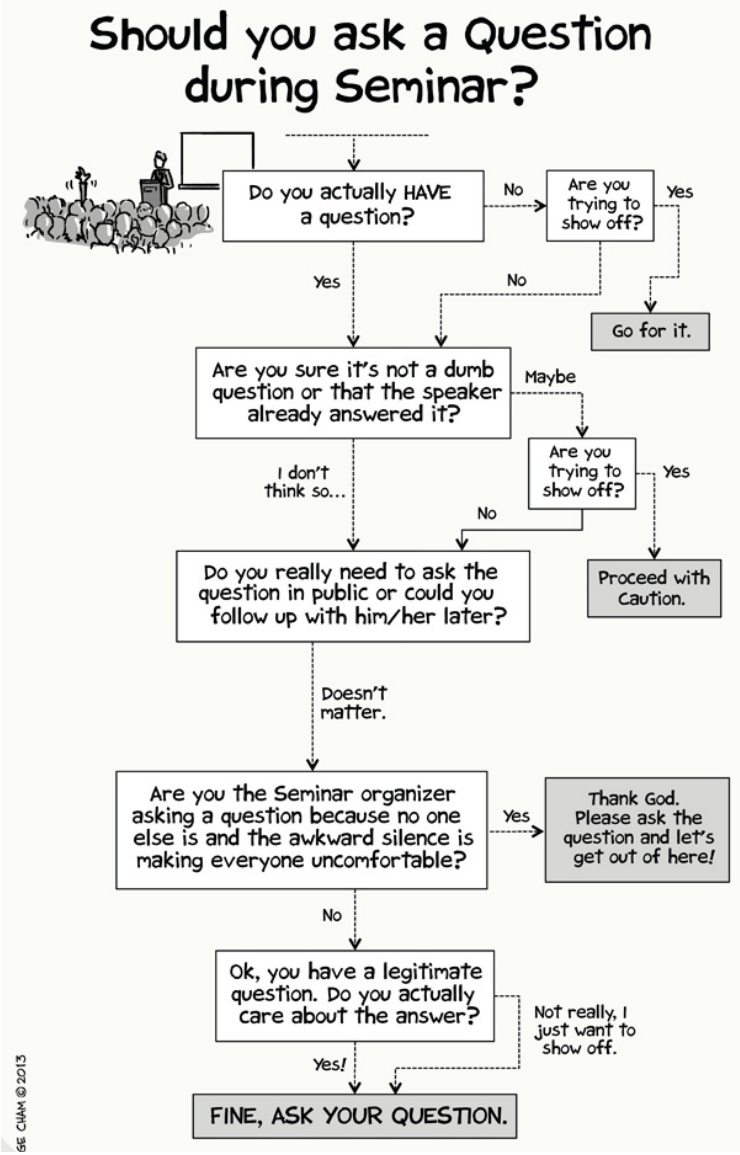
Set up & Requirements:
https://bitly.cx/iHcxS & https://bit.ly/3pyXRii
we will use **Jupyter Notebooks** and Keras / TensorFlow
we will use **Google Colab** → Google Account required

This is a tutorial
→ Please ask questions!



This is a tutorial

→ **I will ask questions to you!**



Deep Learning

- Machine Learning Basics
- Neural Networks
  - Backpropagation, Optimization
  - Activation, Initialization
  - Preprocessing

Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)

A woman is throwing a frisbee in a park. A dog is standing on a hardwood floor. A stop sign is on a road with a mountain in the background.

ArXiv: 1502:03044

KÜNSTLICHE INTELLIGENZ

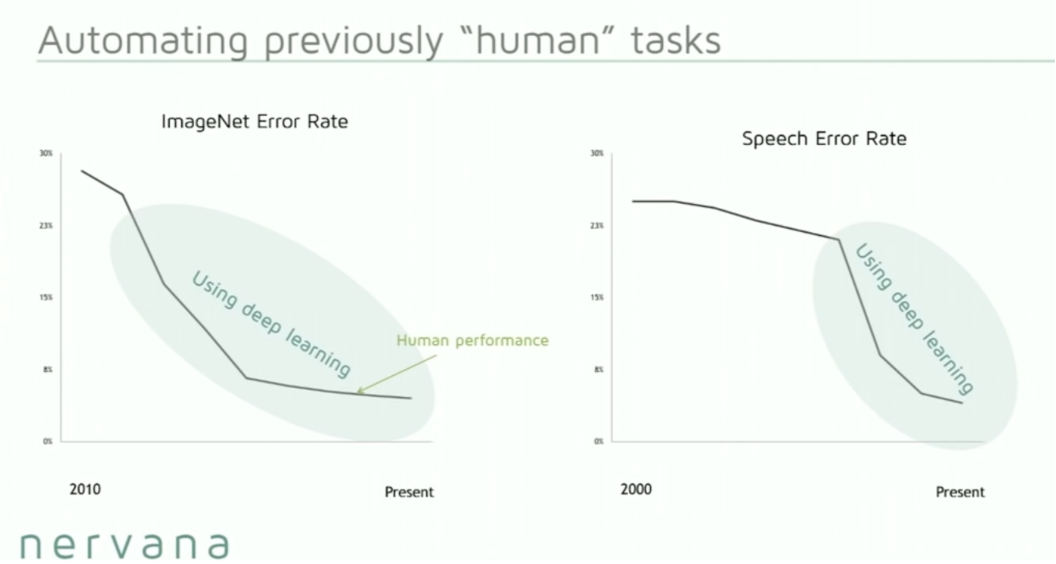
Schlau in zwei Stunden

VON ALEXANDER ARMBRUSTER - AKTUALISIERT AM 27.09.2017 - 11:41

Artificial Intelligence - "The effort to automate intellectual tasks normally performed by humans"

Deep Learning

- Large progress of artificial intelligence due to **Deep Learning**



Example: Caption Generation

Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

ArXiv: 1502:03044

Image Generation

"Albert Einstein using a mobile phone while watching TV"

"A happy carrot on the moon playing chess"

"A penguin flies in the sky and overtakes other birds. Clouds are seen in the background"

Breakthrough in generative machine learning

- generation of realistic images
- image feature local and global coherence
- realistic image super resolution

Large Language Models

Write a poem about the fascination of measuring ultra-high-energy cosmic rays using the Pierre Auger Observatory

In the high expanse where stars hold sway,
A dance of particles in cosmic ballet,
There lies a fascination, a scientific quest,
Measuring rays that leave us all impressed.

Pierre Auger, an observatory high,
A lens to the cosmos, reaching the sky,
Ultra-energy whispers, secrets untold,
In cosmic rays, mysteries unfold.

In the argent fields of celestial might,
Particles cascade, a dazzling light,
Charged messengers from a distant realm,
Pierre Auger's gaze, the cosmic helm.

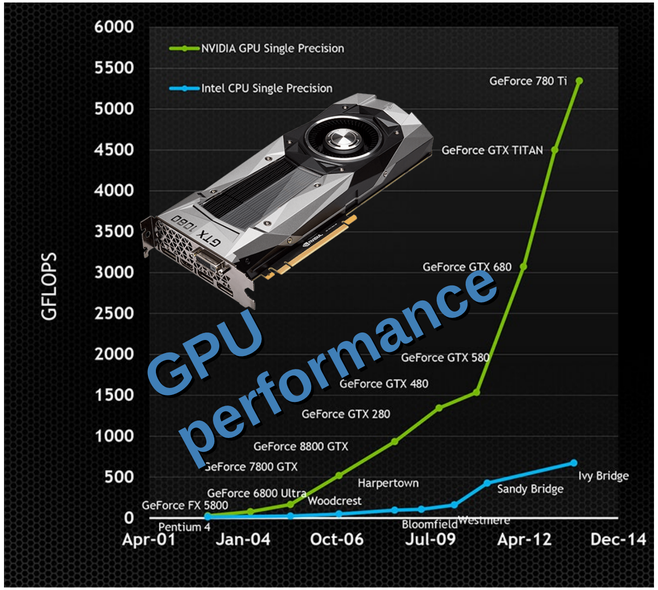
DeepL

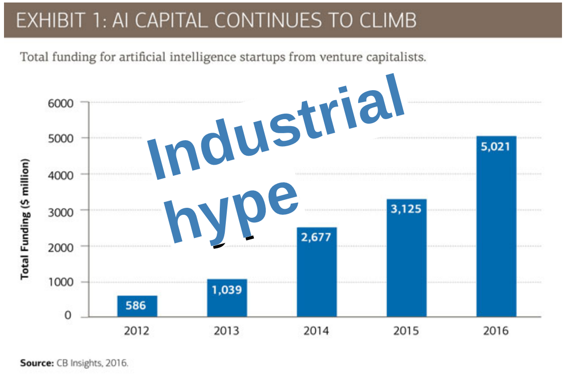
- Breakthroughs in:
  - Machine translation
  - general-purpose language <u>understanding</u> and <u>generation</u>
- Based on transformer networks
- Up to 100s billions of adaptive parameters
- 3M USD for model training
- 400 yrs to train on V100 GPU (price V100: 7,000 €)
- ~1000 MWh for training

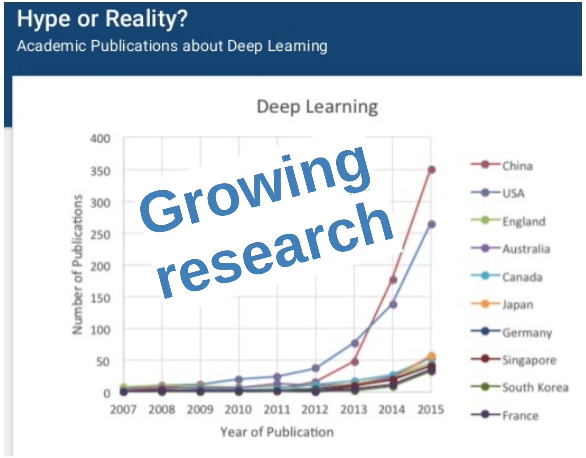
LLMs can be helpful for you daily scientific work (coding, phrasing, searching)

Deep Learning

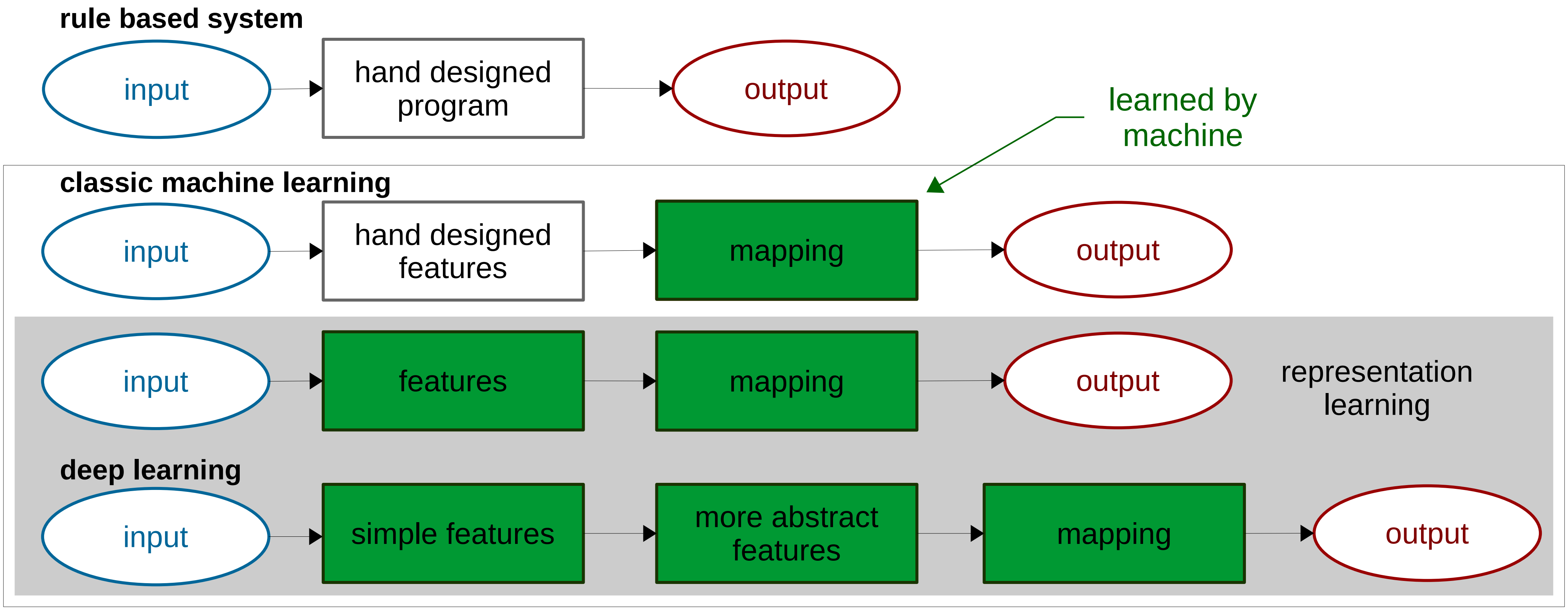
- Every minute:
  - Instagram users post 200,000 photos
  - Twitter users send 350,000 tweets
  - Data on billion scale every day





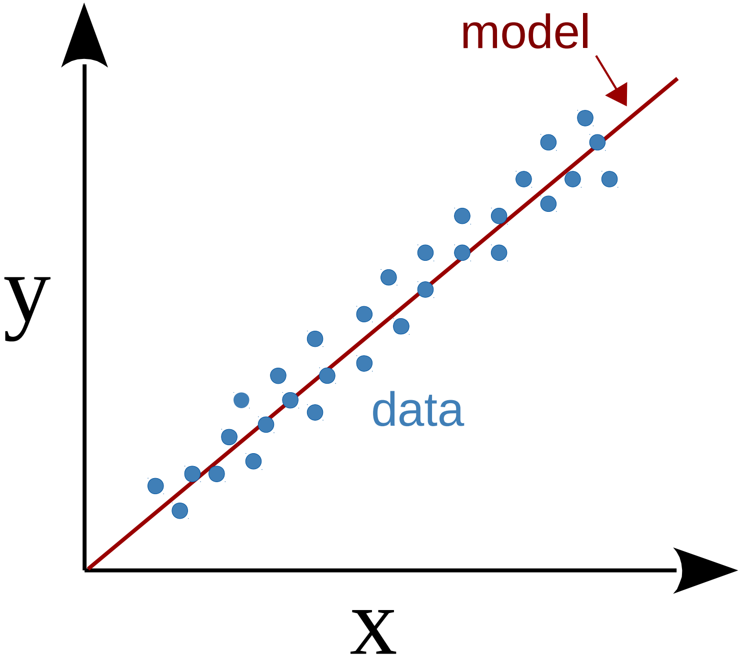


When is it Deep?



"It's deep if it has more than one stage of non-linear feature transformation" *- Y. LeCun*

Machine Learning – Regression



- Data: $\{x_i, y_i\},\ i = 1, ..., N$
- Define model:
  $y_m(x, \theta) = Wx + b$ with free parameters $\theta = (W, b)$
- Define **objective function** (loss/cost)

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} [y_m(x_i, \theta) - y_i]^2$$

- Train model (minimize objective) $\hat{\theta} = argmin[J(\theta)]$
  - Optimize set of free parameters $\theta = (W, b)$
    eg. use gradient descent

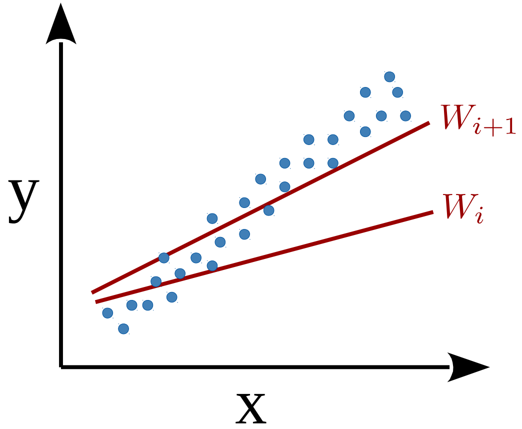
Gradient Descent

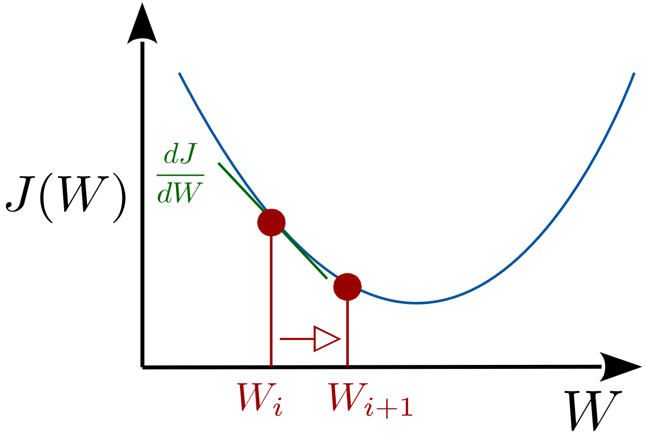
- Minimize objective function $J(\theta)$ by updating θ in **opposite** direction of gradient iteratively

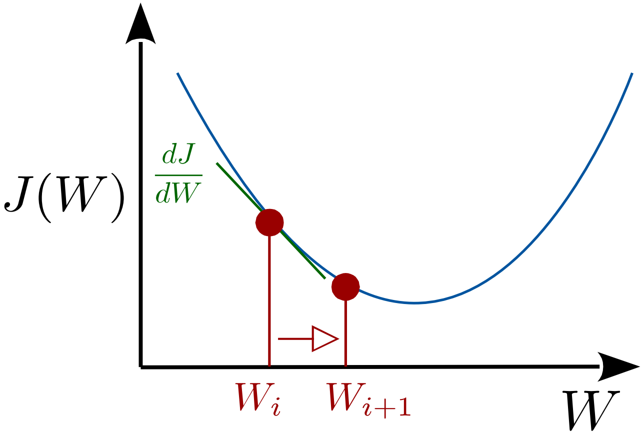
gradient:	$dJ/d\theta$
stepsize:	α

$$\tilde{\theta} \rightarrow \theta - \alpha \frac{dJ}{d\theta}$$

- Example: linear regression with mean squared error







Is the loss surface always parabolic?

(a) Yes, this is why the MSE is so nice!

(b) No, only when using the parabolic MSE los $(x-y)^2$

(c) No, only in the special case of linear regression!

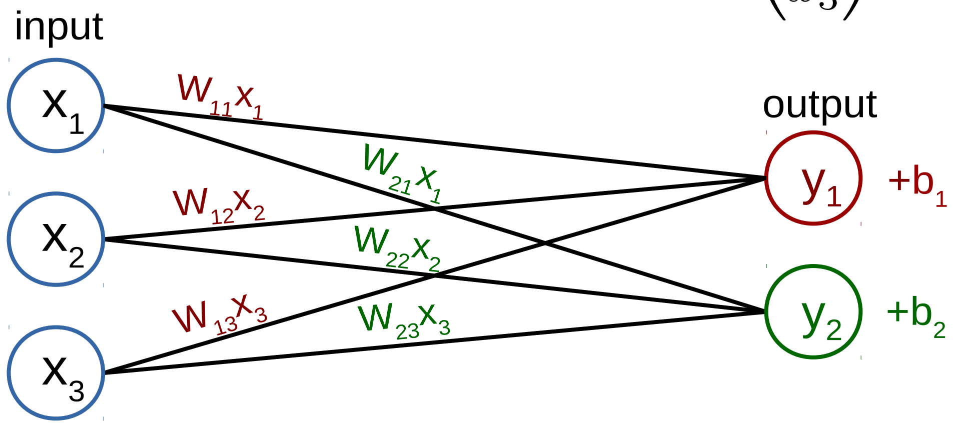
Multidimensional Linear Models

- Predict multiple outputs $\mathbf{y} = (y_1, ..., y_n)$ from multiple inputs $\mathbf{x} = (x_1, ..., x_n)$ using linear function $\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$

Note: We define linear = affine in this course

- Example: $x \in \mathbb{R}^3,\ y \in \mathbb{R}^2$

$$\begin{pmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \end{pmatrix} \times \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$



Non-Linear Network Models

$\mathbf{Wx} + \mathbf{b}$ only describes linear models

- Use network with several linear layers:

$$h' = W^{(1)}x + b^{(1)}$$

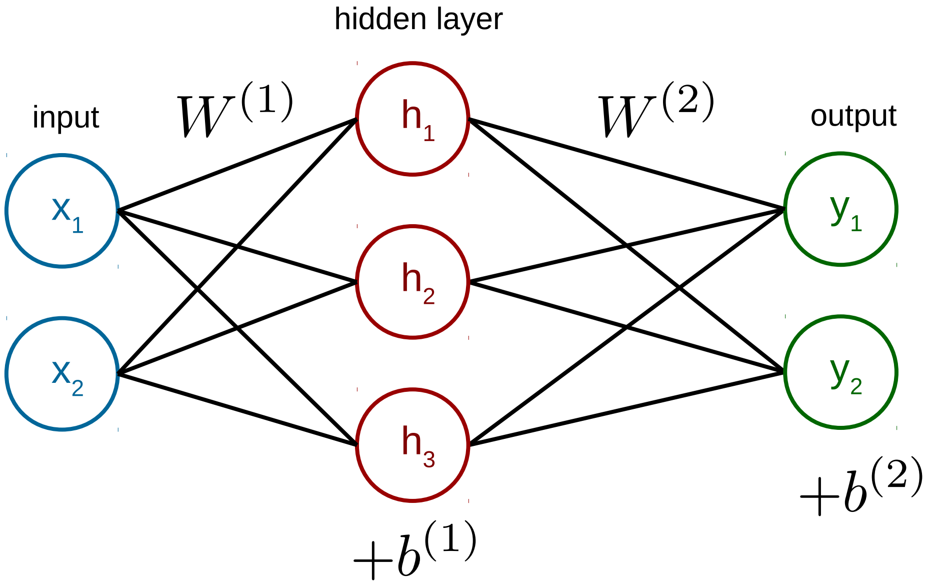
$$y = W^{(2)}h' + b^{(2)}$$

- Model is still linear!

$$y = W^{(2)}\left(W^{(1)}x + b^{(1)}\right) + b^{(2)}$$

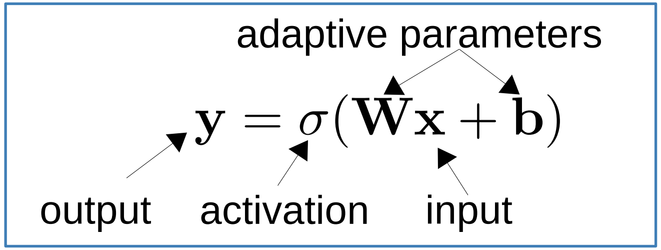
$$y = \underbrace{W^{(2)}W^{(1)}}_{W}x + \underbrace{W^{(2)}b^{(1)} + b^{(2)}}_{b}$$

➢ Solution: Apply non-linear activation σ to each element $\longrightarrow h = \sigma(h') = \sigma(Wx + b)$



Activation Functions

- Using an activation function the layer becomes a non linear mapping
  - Allows for stacking several layers



Examples

- **Rectified Linear Unit**

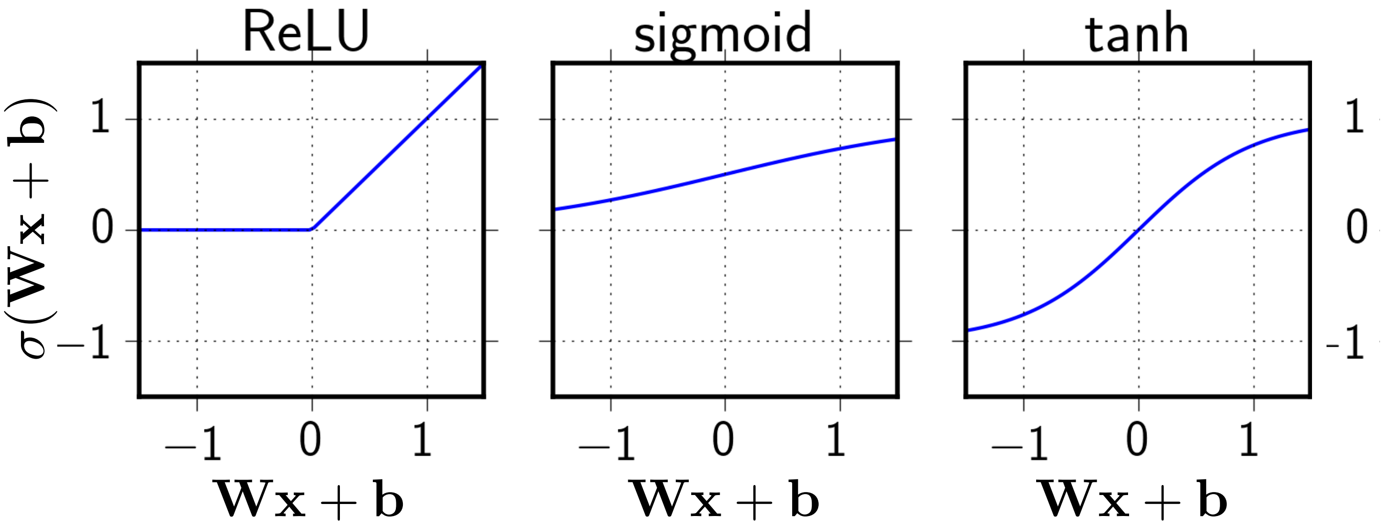
$$\sigma(x) = \max(0, x)$$

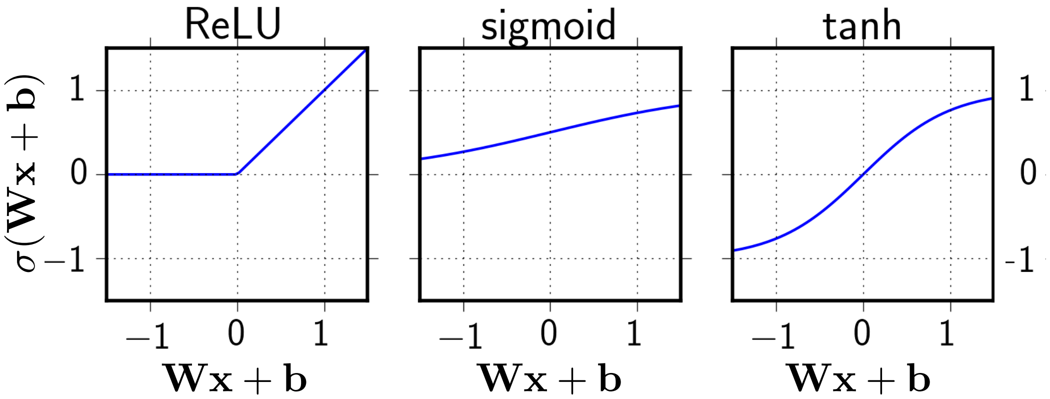
- **Sigmoid**

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

- **Hyperbolic tangent**

$$\sigma(x) = \frac{e^{+2x} - 1}{e^{-2x} + 1}$$





What is a nice activation function?

(a) ReLU → since it's so simple and has a simple and constant gradient

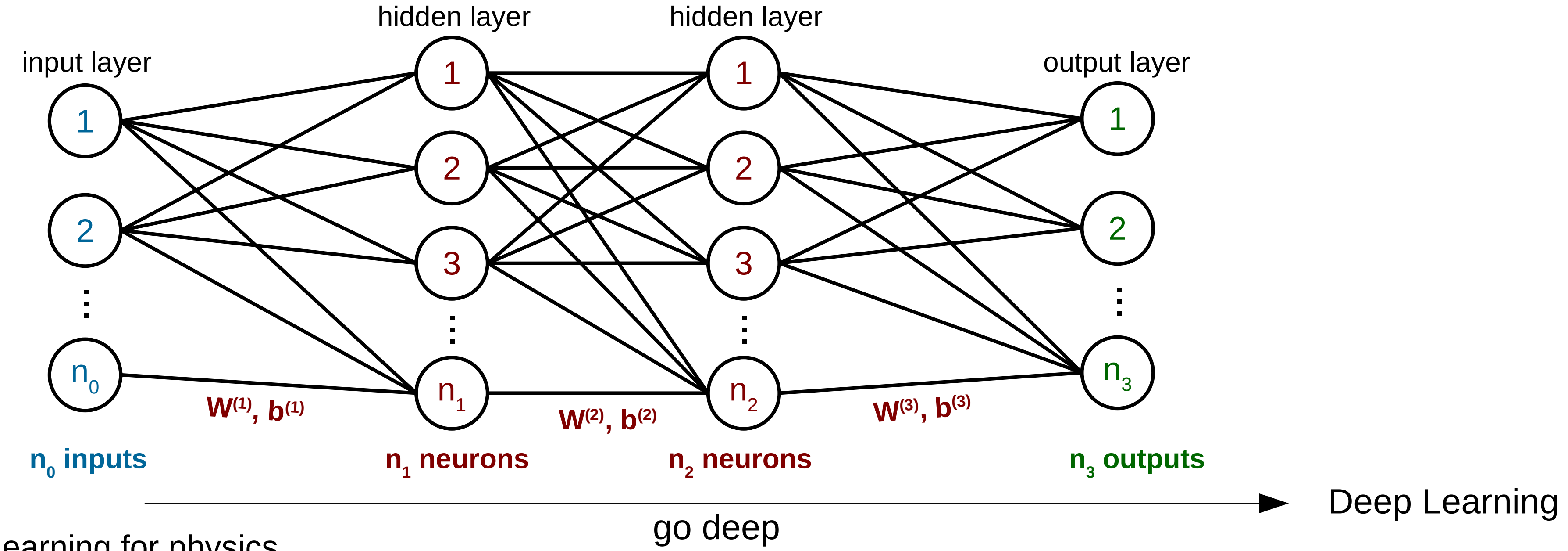
(b) Sigmoid → it's very complex and inspired by biology

(c) Tanh → it is complex and also permits negative values

Neural Networks

Basic unit $\sigma(Wx + b)$ is called **node/neuron** (analogy to neuroscience)

- Strength of connections between neurons is specified by **weight matrix** W
- **Width:** number of neurons per layer
- **Depth:** number of layers holding weights (do not count input layer)

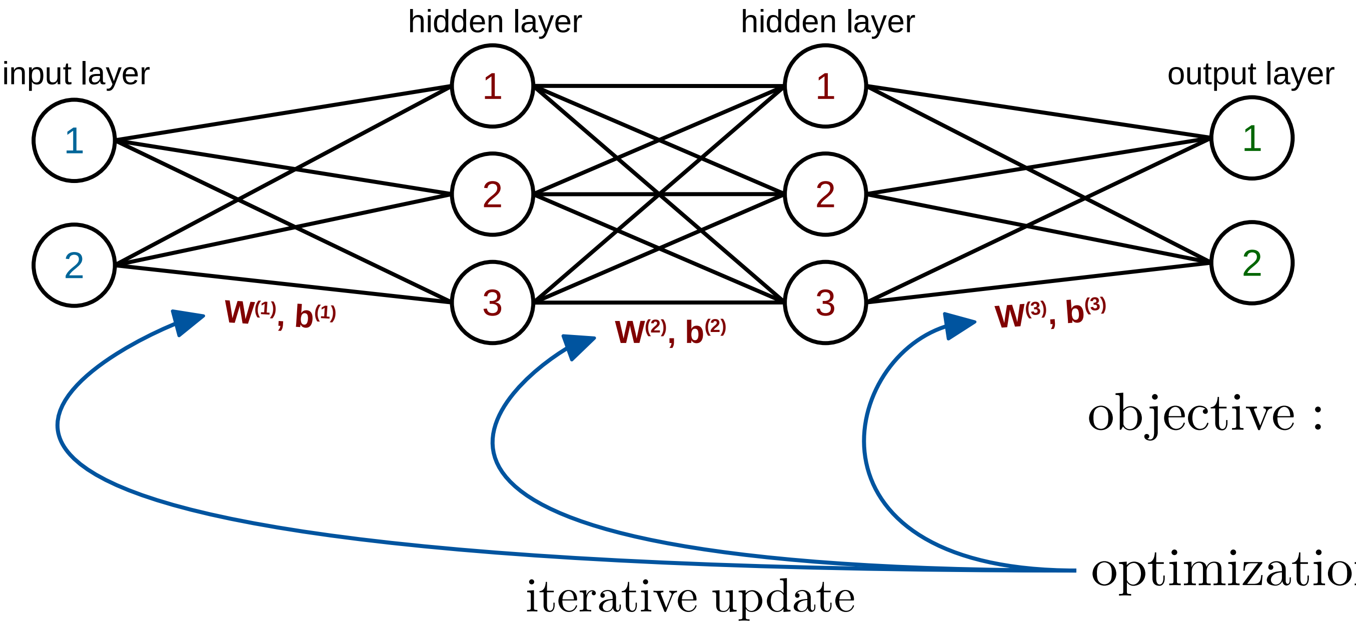


Feature Learning

Feature Hierarchy: each new layer extract more abstract information of the data.

Probabilistic Mapping: learns to combine the extracted features

Train model (to find $\theta = \{W_i, b_i\}$ that minimizes objective) is automatic process.



$$\text{objective}: \quad J(\theta) = \sum_i \left[y_m(x_i, \theta) - y_i\right]^2$$

$$\text{optimization}: \quad \frac{dJ}{d\theta} \to 0$$

Initialization

- **Weights need different (random) initial values → symmetry breaking**
- Scale of weights very important
  - Too large → exploding signals & gradients
  - Too small → vanishing signals & gradients

  No learning!

- For forward pass in each layer:

$$Var[x_l] = 1$$

- For Backward pass in each layer:

$$Var[\Delta x_l] = 1$$

➢ Depends from activation function and number of in and outgoing nodes

$$Var[W] = \frac{2}{n_{\text{in}} + n_{\text{out}}}$$

→ For tanh

Glorot, Bengio

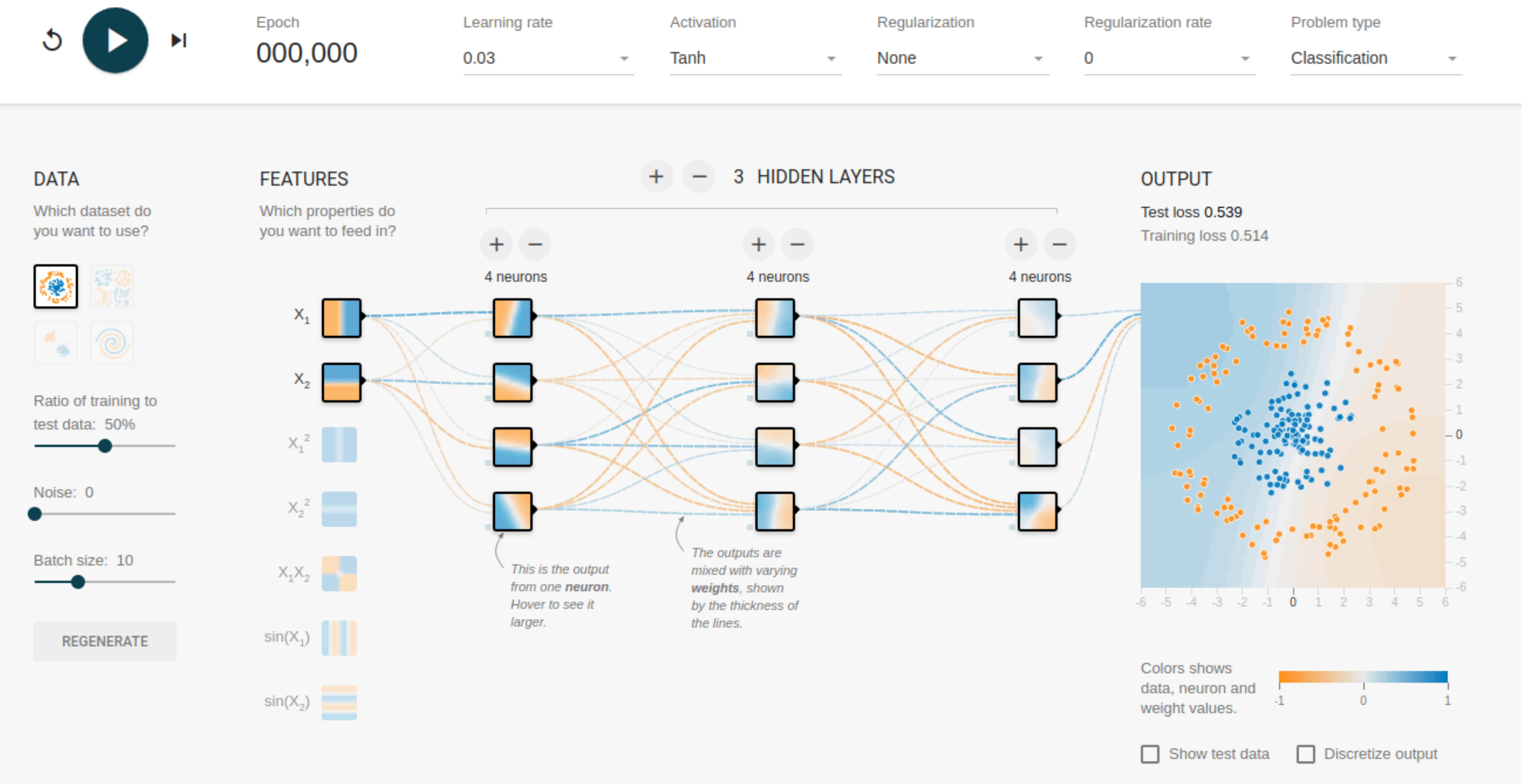
$$Var[W] = \frac{2}{n_{\text{in}}}$$

→ For ReLU

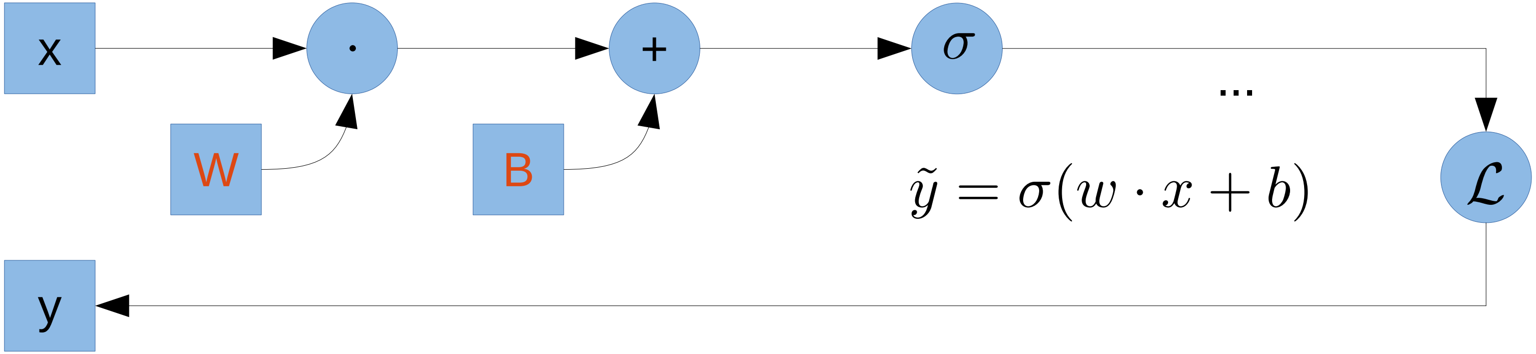
He et al.

- Can be sampled from Gaussian or uniform distribution (Var. scaled by factor of 3)

Example Training



Backpropagation



- Network is series of simple operations (linear mappings/activations/loss ...)
- Use chain rule to evaluate gradient for each parameter → **Backpropagation**

label, prediction

for ReLU simply 0 or 1 → $\sigma'(\tilde{x})|_{\tilde{x}=(w \cdot x + b)}$

input

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial (y - \tilde{y})^2}{\partial w} = \frac{\partial (y - \tilde{y})^2}{\partial \tilde{y}} \cdot \frac{\partial \tilde{y}}{\partial w} = -2 \cdot (y - \tilde{y}) \cdot \frac{\partial \sigma(w \cdot x + b)}{\partial w} \cdot x$$

adaptive parameter

deeper models: x would be output of previous layer

➙ **no need to evaluate full gradient, later part already estimated**

➙ **gradient is "propagated backwards"**

Gradient Decent: Learning Rate

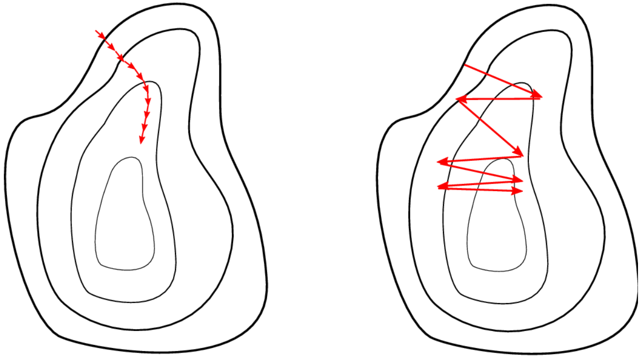
- Learning rate α determines speed of training
- High rate
  - poor convergence behavior or none at all
- Small rate
  - Very slow training or none at all
- Typical learning rate $\alpha = 10^{-3}$

$$\theta \rightarrow \theta - \alpha \frac{dJ}{d\theta}$$

Learning rate

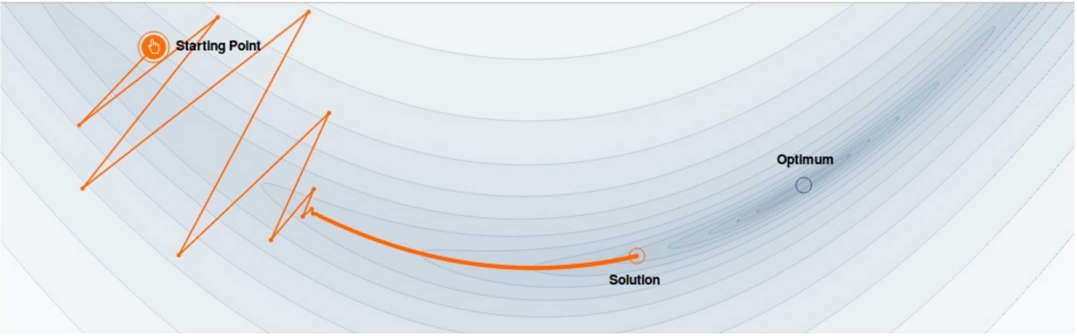
Advanced

- Reduce learning rate when loss stops decreasing
  - increase sensitivity to smaller scales



α too small α too large

Stochastic Gradient Descent - SGD



Why Momentum Really Works, Distill

- Use small subset (mini batch) of dataset for calculating the gradient
  - 1 **epoch** = full pass through training data set
  - Reduces computational effort
  - More updates per epoch → speeds up convergence
  - Stochastic behavior → improve generalization performance
- **Batch size** is hyperparameter and mostly in order of ~32

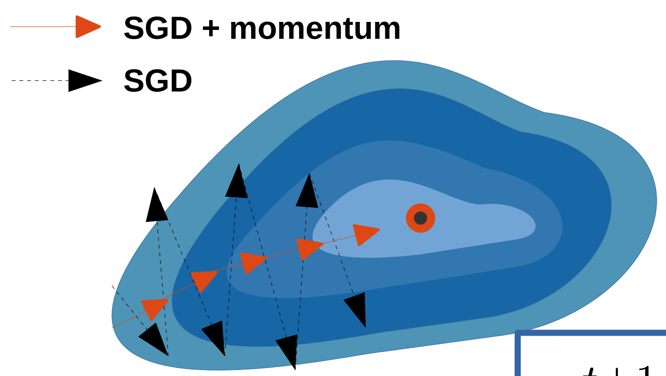
Advanced Optimizer

Momentum: Use past gradients (velocity)

- Faster convergence by **damping oscillations** and increasing the step size for more informative gradients

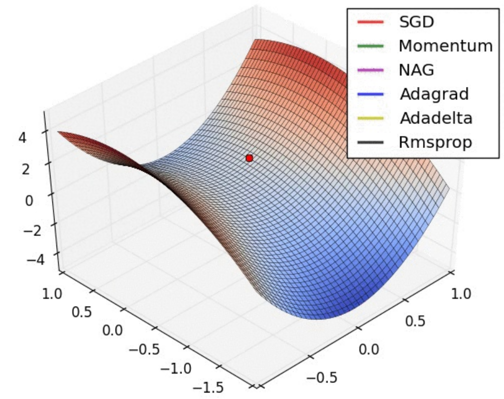
Adaptive learning rate: Scaling using past gradients (Adagrad, **Adam**, Adadelta...)

- Use adaptive learning rates for each parameter



$$z^{t+1} = \beta z^t + \nabla f(\theta^t)$$
$$\theta^{t+1} = \theta^t - \alpha \cdot z^{t+1}$$

Convergence behavior of optimizers



Sebastian Ruder: http://ruder.io/optimizing-gradient-descent/

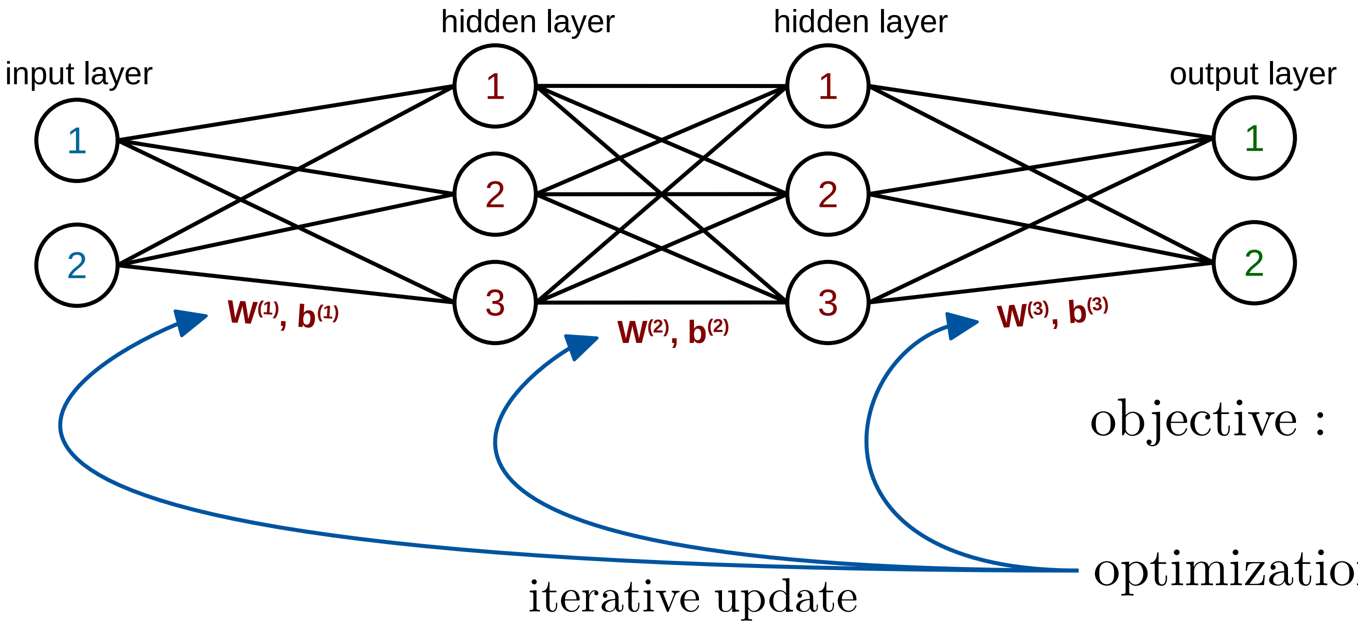
https://distill.pub/2017/momentum/

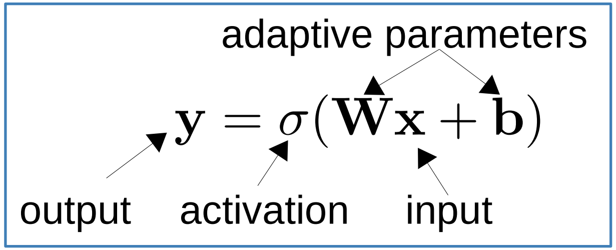
Deep Neural Networks

Feature Hierarchy: each new layer extract more abstract information of the data.

Probabilistic Mapping: learns to combine the extracted features

Train model (to find $\theta = \{W_i, b_i\}$ that minimizes objective) is automatic process.

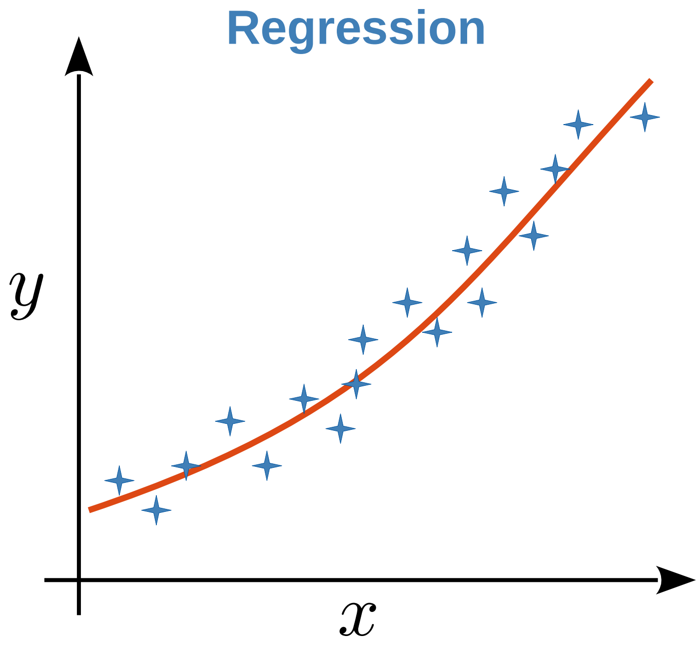


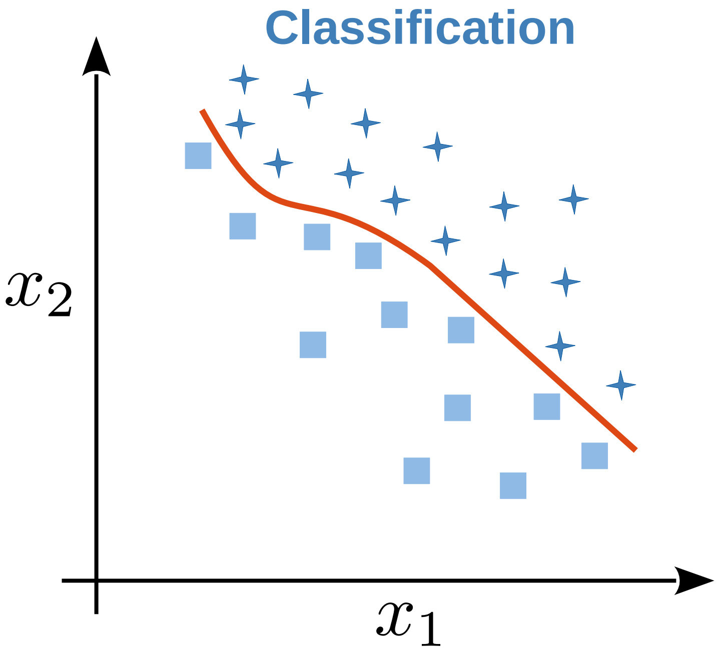


$$\text{objective}: \quad J(\theta) = \sum_i [y_m(x_i, \theta) - y_i]^2$$

$$\text{optimization}: \quad \frac{dJ}{d\theta} \to 0 \qquad \tilde{\theta} \to \theta - \alpha \frac{dJ}{d\theta}$$

Machine Learning Tasks

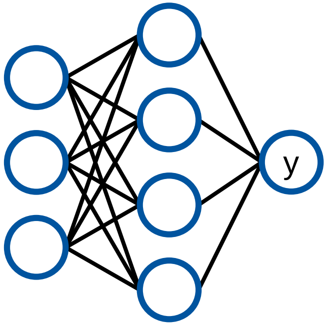




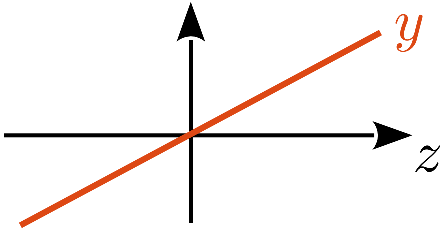
- Regression: Predict continuous label y
- Classification: Separate into different classes (cats, dogs, airplanes, …)
- Can sometimes convert to the other

Classification vs. Regression

Regression



Linear

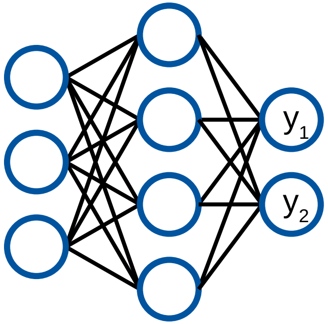


no activation function

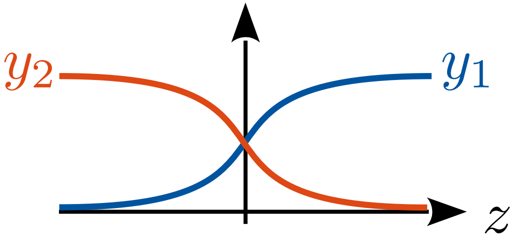
Minimize mean-squared-error

$$J(\theta) = \frac{1}{n} \sum_i [y_i - y_m(x_i)]^2$$

Classification



Softmax



$$y_j(z) = \frac{e^{z_j}}{\sum_i e^{z_i}}$$

Minimize cross entropy

$$J(\theta) = -\frac{1}{n} \sum_i y_i \log[y_m(x_i)]$$

TensorFlow Playground - 15 Minutes

Checkerboard task

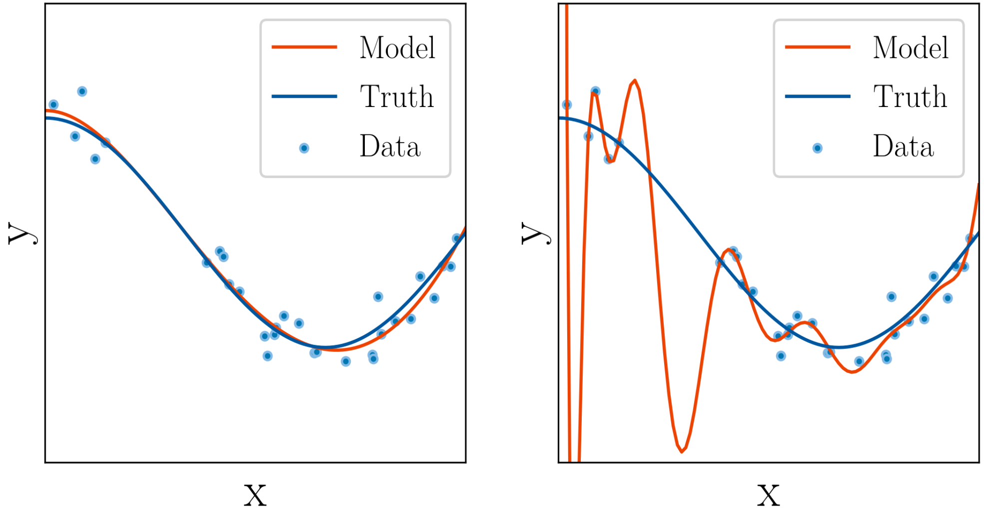
- Choose the Checkerboard data set (XOR)
- What do you observe when changing the activation function?
- What do you see when inspecting the features of deeper layers?
- Choose the ReLU activation:
  - What is the **minimum** number of nodes / layers needed to solve the task?

Open the example at:

https://playground.tensorflow.org/

Bonus: Which extra feature is most useful?

Generalization

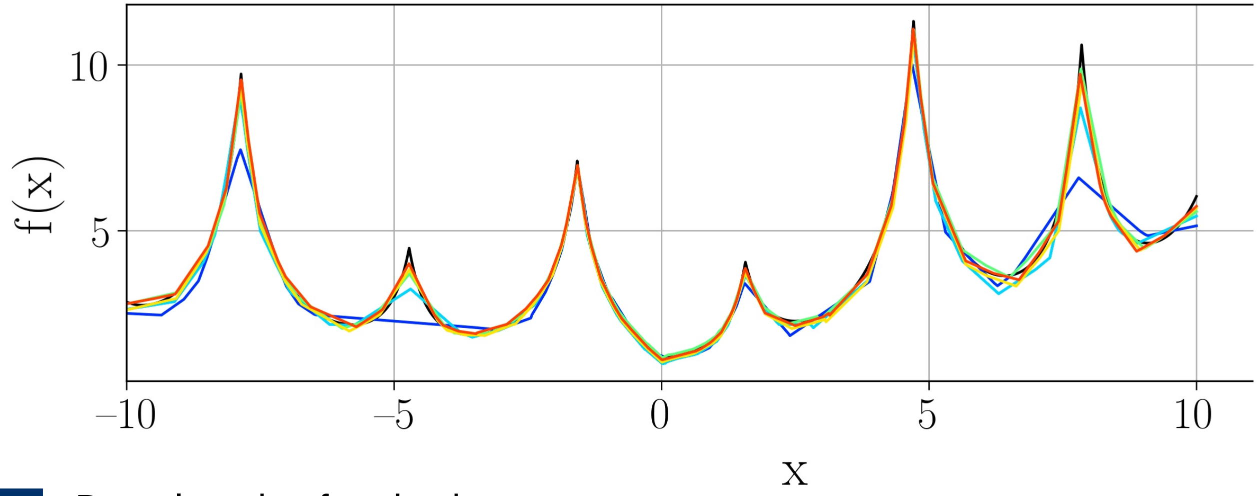


I. Training, Validation, Testing
II. Under- and Overfitting
III. Regularization

Universal Approximation Theorem

“A feed-forward network with a linear output and at least ***one hidden layer*** *with a finite number of nodes can (in theory) approximate any reasonable function to arbitrary precision.”*

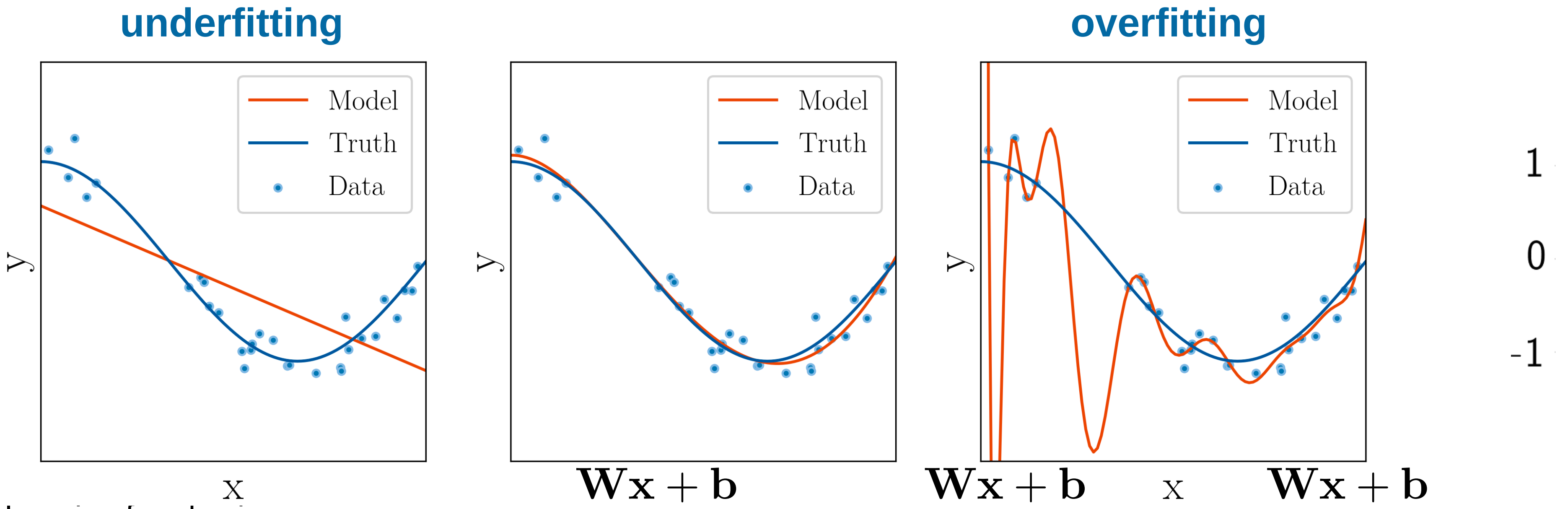
- Network design considerations → feature engineering, network architecture
  - Shallow networks often show bad performance → train deep models!



- Fit complicated function
  - Use neural network
  - 2 hidden layers a 30 nodes

Under- and Overfitting

- Challenging to find a good network design
- Under-complex models show bad performance
- complex models are prone to overfitting
  - Model memorizes training data under loss of generalization performance



Generalization & Validation

A complex network can learn any function, how can we monitor overfitting?

Generalization

Unknown true distribution $p_{true}(x, y)$ from which data is drawn.

Trained model $y_m(x)$ provides prediction based on this limited set

- How good is the model when faced with new data?

Validation

Estimate generalization error on data not used during training.

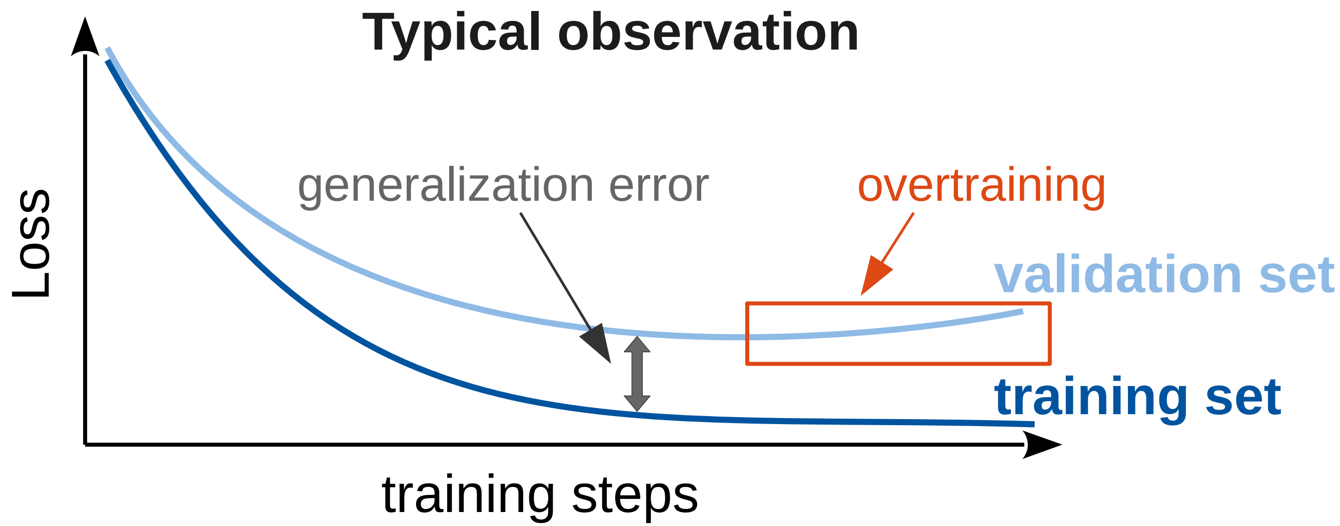
Split data into:

- **Training set:** to train the network
- **Validation set:** to monitor and tune the training (training of hyperparameter)
- **Test set:** to estimate final performance. Use only **once!**

Why can’t we use the validation data set for testing?

Under- and Overtraining

- During training monitor the loss separately for training and validation set



Training loss:
- decreases

Validation loss:
- is higher than training loss → **generalization gap**
- has a minimum → **overtraining**

What is a clear sign of overtraining?

(a) Some large weights (they contribute most)

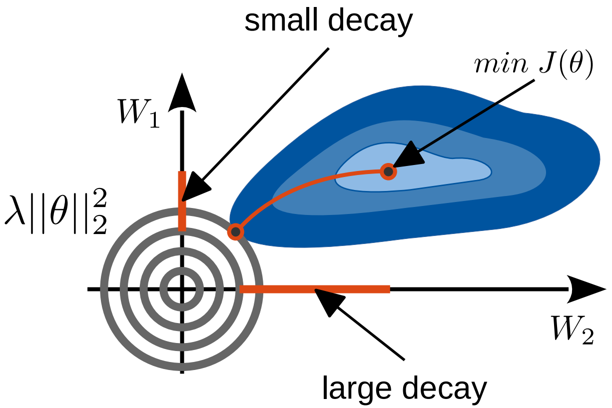
(b) Many average weights (all do the same)

(c) Many small values (DNN learns almost nothing)

Parameter Norm Penalties

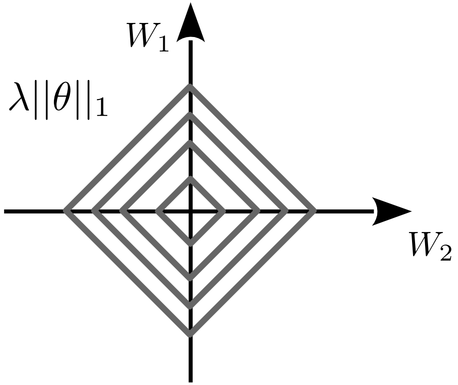
L² norm: (**weight decay**) $\lambda||\theta||_2^2 = \lambda(\theta_1^2 + \theta_2^2 + ...)$

- Contribution to loss dominated by largest weights
- Decay of weights which not contribute much to the reduction of the objective $J(\theta)$



L¹ norm: (**lasso**) $\lambda||\theta||_1 = \lambda(|\theta_1| + |\theta_2| + ...)$

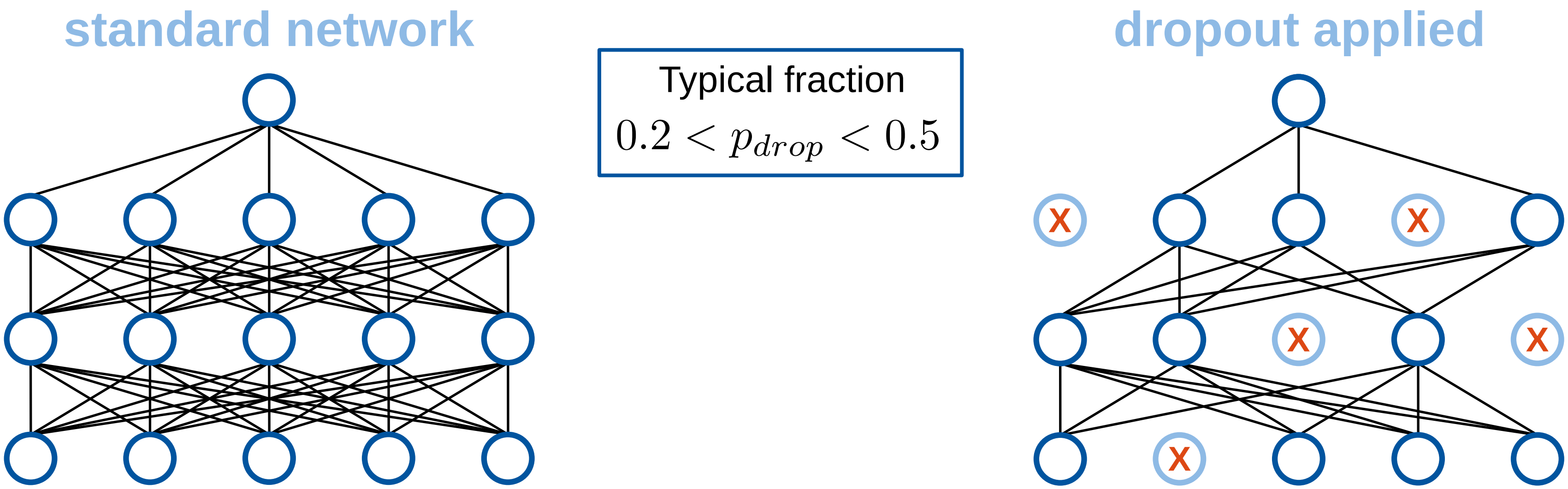
- Constant shrinking of parameters
- Allows for sparse network (feature selection mechanism)



ElasticNet: Combination of L¹ and L² norm

Dropout

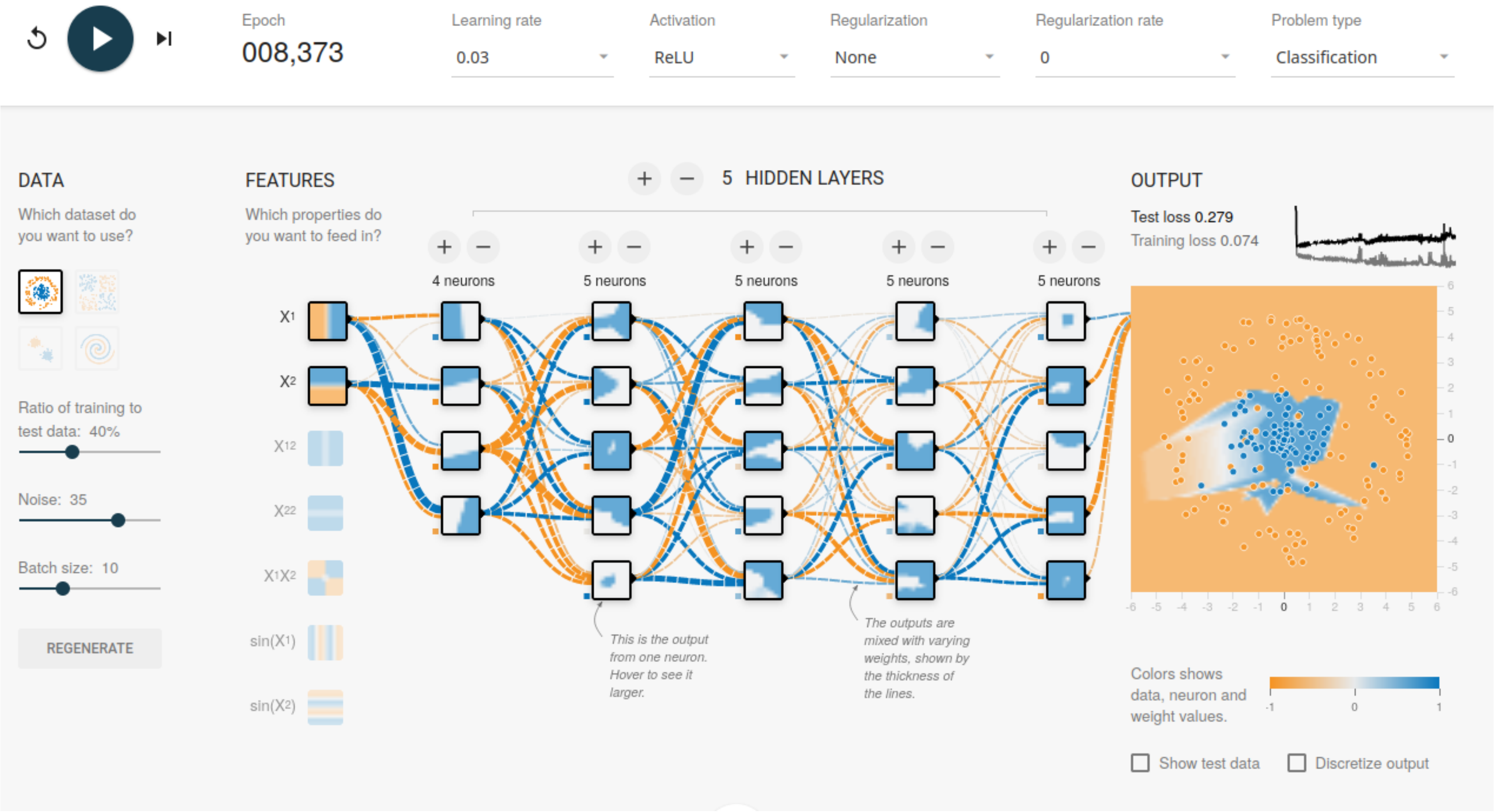
Randomly turn of fraction p_{drop} of neurons in each training step



Typical fraction
$0.2 < p_{drop} < 0.5$

- Adds noise to process of feature extraction
- Force network to train redundant representations
- During validation and test: no dropout applied → large ensemble of “submodels”

Overtraining



TensorFlow Playground - 15 Minutes

Checkerboard task

- Choose the Checkerboard data set (XOR)
- Set noise to 50%, choose a deep network and train for 1000 epochs
- Apply L2 regularization to reduce overfitting. Try low and high regularization rates. What do you observe?
- Compare the effects of L1 and L2 regularization.

Open the example at:

https://playground.tensorflow.org/

TensorFlow Playground - 15 Minutes

Solution

Clarifying frequent misunderstandings

- **Use of activation functions -** layer without activation is usually meaningless
  - sigmoid **only** @ last layer in classification / regression @ last layer **no** activation
- **Universal approximation theorem is only a theoretic statement**
  - even such models exists → you have to find its design & **train** it → not easy!
- **Test and validation data are different**
  - validation: tune your DNN, e.g. train 10 DNNs & compare, monitor overtraining
  - test: check after you decide for one of the 10 models → ONCE!
- **Training networks is not random** → extract features out of patterns in data
  - retraining gives slightly different DNN → its feature sensitive to same patterns!
- **DNNs are not the holy grail** → simple fits can outperform DNNs
  - lots of data needed, challenge has to be complex and multi-dimensional

Jonas Glombitza
jonas.glombitza@fau.de

ECAP ERLANGEN CENTRE FOR ASTROPARTICLE PHYSICS

FAU

Deep Learning for Physics Research

Exercise class:
- fully-connected networks
- convolutional neural networks

Set up & Requirements:
https://bitly.cx/iHcxS & https://bit.ly/3pyXRii
we will use **Jupyter Notebooks** and Keras / TensorFlow
we will use **Google Colab** → Google Account required

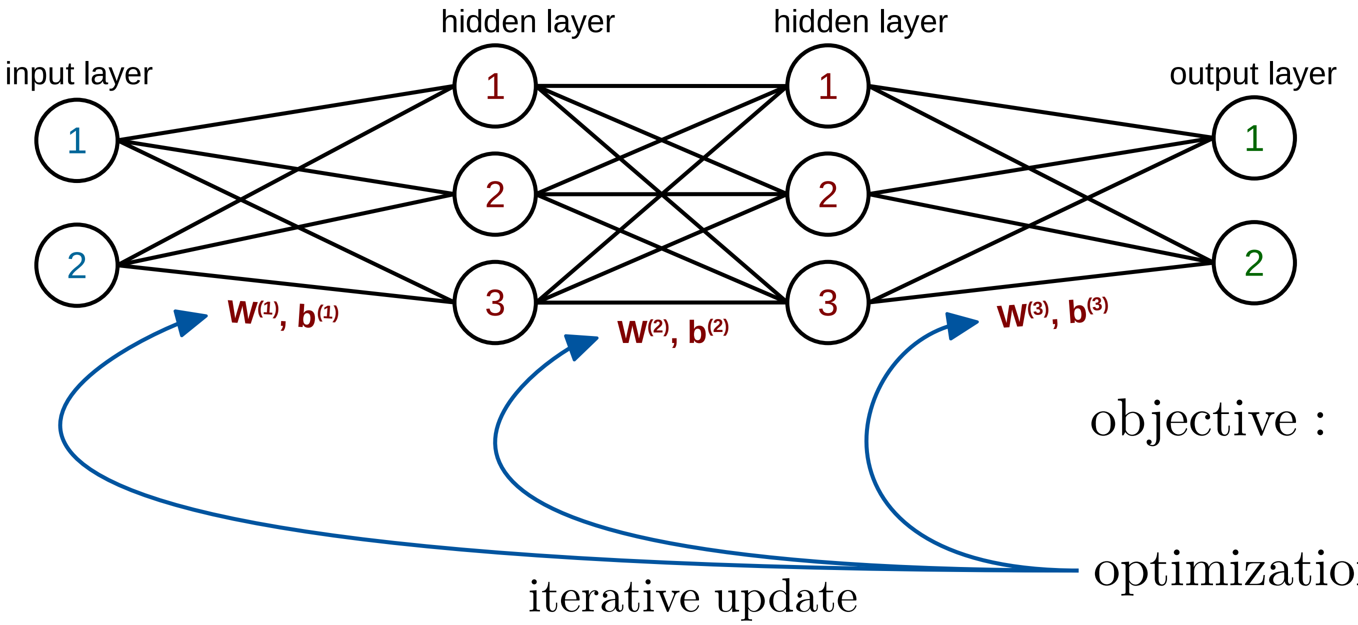
https://github.com/DeepLearningForPhysicsResearchBook/deep-learning-physics

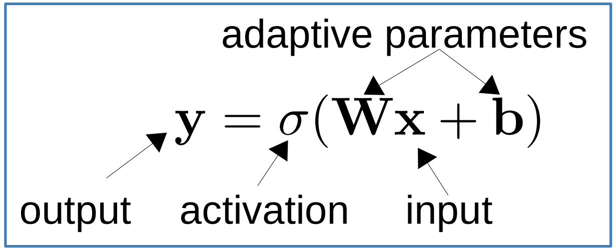
Deep Neural Networks

Feature Hierarchy: each new layer extract more abstract information of the data.

Probabilistic Mapping: learns to combine the extracted features

Train model (to find $\theta = \{W_i, b_i\}$ that minimizes objective) is automatic process.





$$\text{objective}: \quad J(\theta) = \sum_i \left[y_m(x_i, \theta) - y_i\right]^2$$

$$\text{optimization}: \quad \frac{dJ}{d\theta} \to 0$$

$$\tilde{\theta} \to \theta - \alpha \frac{dJ}{d\theta}$$