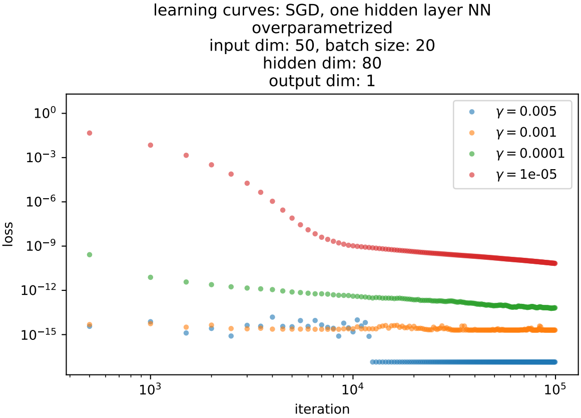
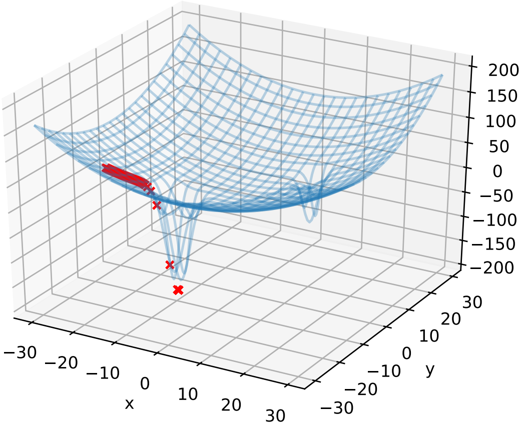
Supervised learning: neural networks




200
150
100
50
0
−50
−100
−150
−200
−30
−20
−10
0
10
20
30
x
30
20
10
0
−10
−20
−30
y

Vectors, matrices, neural nets

Derivation, gradient descent, backpropagation

Training and visualizing some neural nets

Other techniques

MNIST

Overparametrization and underparametrization

Ressources

- https://www.deeplearningbook.org/
- https://course.fast.ai/
- https://d2l.ai/
- https://mlelarge.github.io/dataflowr-web/dldiy_ens.html
- https://playground.tensorflow.org/
- http://www.jzliu.net/blog/simple-python-library-visualize-neural-network/

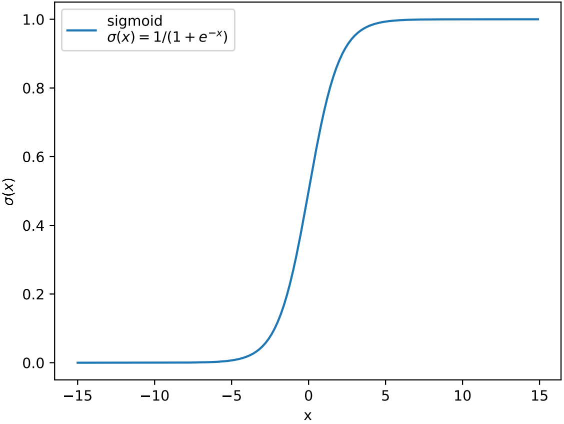
Elementary neuron

A **neuron** is a **mapping** from a multidimensional input $x = (x_1, ..., x_d)$ to a real number.
This function depends on **parameters** called the **weights** $w = (w_1, ..., w_d)$ and the **intercept** b (omitted in the following slides to simplify the notations).

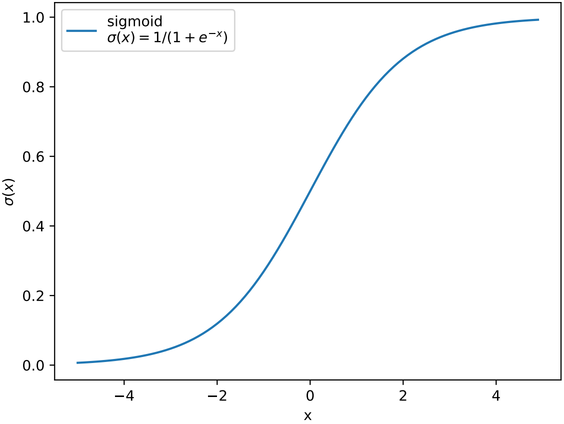
$$f(x) = \sigma(\sum_{i=1}^{d} x_i w_i) \tag{1}$$

Where σ is a non linear function, for instance a **sigmoid**.

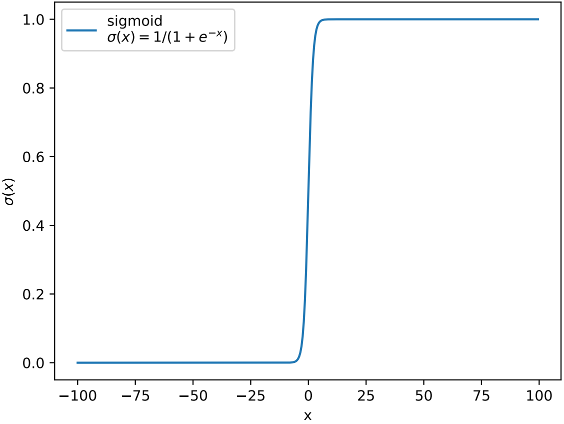
Sigmoid function



Sigmoid function



Sigmoid function



Notation with vectors

The sum $\sum_{i=1}^{d} x_i w_i$ can also be written this way :

$$xw^T \tag{2}$$

This means a **product** of **two matrices** (a vector is also a matrix : it is just a matrix with only one line or only one column):

- $x = (x_1, ..., x_d)$
- $$w^T = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

Matrices

A matrix is an array used to store data. It has **lines** and **columns**

$$A = \begin{pmatrix} 1 & 4 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

A_{ij} means the element at line i and column j.

- $A_{12} = 4$
- $A_{31} = 0$
- $A_{33} = 1$

Matricial multiplication

- If matrix A has p columns and matrix B has p lines, the product AB of the two matrices is defined as:

$$AB_{ij} = \sum_{k=1}^{p} A_{ik} B_{kj} \tag{3}$$

- `https://en.wikipedia.org/wiki/Matrix_multiplication`
- It is often more convenient and compact to write computations as operations on vectors and matrices when possible.

Examples

$$\begin{pmatrix} 1 & 4 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = ?$$

Examples

$$\begin{pmatrix} 1 & 4 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Examples

$$\begin{pmatrix} 1 & 4 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = ?$$

Examples

$$\begin{pmatrix}1 & 4 & 1\\0 & 1 & 2\\0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{pmatrix}=\begin{pmatrix}1 & 4 & 1\\0 & 1 & 2\\0 & 0 & 1\end{pmatrix}$$

Examples

$$\begin{pmatrix} 1 & 4 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 0 & 0 \end{pmatrix} = ?$$

Examples

$$\begin{pmatrix}1 & 4 & 1\\ 0 & 1 & 2\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\ 0 & 2 & 0\\ 1 & 0 & 0\end{pmatrix} = \begin{pmatrix}2 & 8 & 0\\ 2 & 2 & 0\\ 1 & 0 & 0\end{pmatrix}$$

Examples

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

What is A^n ?

Examples

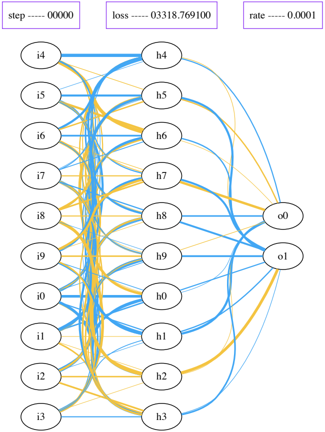
- $x = (x_1, ..., x_d)$
- $w = (w_1, ..., w_d)$

$$xw^T = \sum_{i=1}^{d} x_i w_i \tag{4}$$

Neural networks

- $\sigma(xw^T)$ allows us to compute the output of a **single neuron**
- But we will often have **several neurons** outputting a result.
- These neurons are organized in a network called neural network.

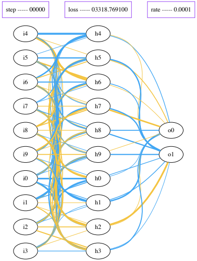
- Let $w = [w_{ij}]$ be the matrices of weights between neuron i of the left layer and neuron j of the middle layer.



Matrices and neural networks

- let $w = [w_{ij}]$ be the matrices of weights between neuron i of the left layer and neuron j of the middle layer.
- the input b_j of the hidden layer as a function of the outputs x_k of the input layer writes:

$$b_j = \sum_{k=1}^{d} x_k w_{kj} \tag{5}$$

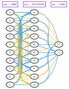


- let $w = [w_{ij}]$ be the matrices of weights between neuron i of the left layer and neuron j of the middle layer.

$$b_j = \sum_{k=1}^{d} x_k w_{kj} \tag{6}$$

- With $b = (b_1, ..., b_m)$ and $x = (x_1, .., x_d)$, we have

$$b = xw \tag{7}$$



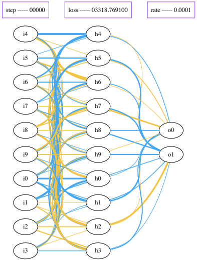
- The same notation generalizes to an arbitraty number of input samples $x_i, i \in [1, n]$!

Layers

- We now know how to compute the input b of a layer as a function of the output x of the previous layer

$$b = xw \tag{8}$$

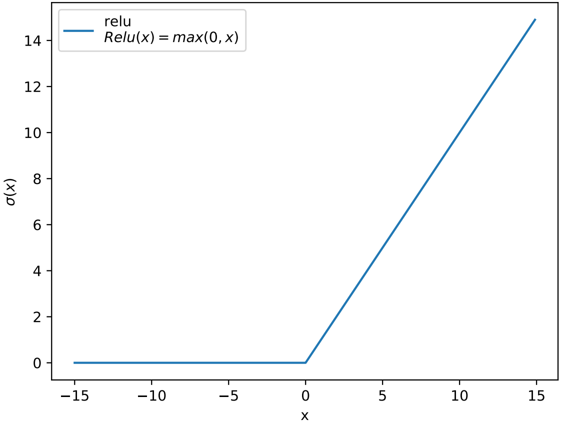
- Applying this rule **and** the non linearity σ, we can coompute the **forward propagation** of a neural network.



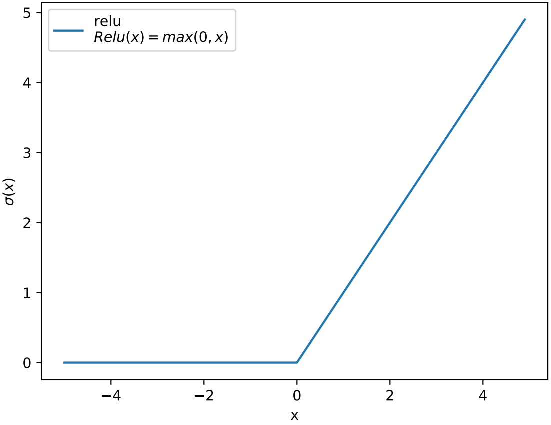
Layers and forward propagation

- We will use the **numpy** library to do so.
- We will use de ReLu as the non linearity.

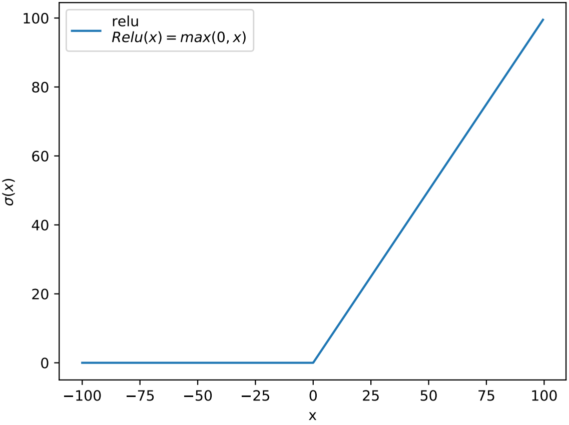
Relu function



Relu function

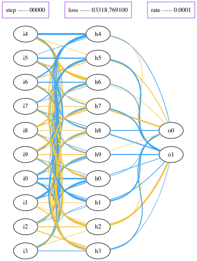


Relu function



Optimization

- The parameters are the **weights** w_1 and w_2.
- In our examples we will use a network with three layers : input - hidden - output.



Gradients in a neural network

- Example of a gradient

$$\frac{\partial L}{\partial w1} = h^T 2(y_{\text{predicted}} - y_{truth}) \tag{9}$$

(where *h* is the output of the relu)

Backpropagation

- By repeating the same process we can also compute the gradient with respect to $w2$.
- This is called **backpropagation**.
- Knowing the gradient, we can **update the network parameters**.

Libs

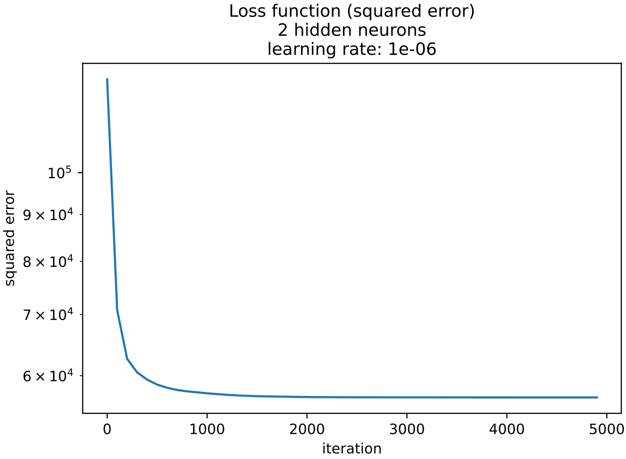
- We will need **numpy**
- **pygraphviz**

Learning toy data

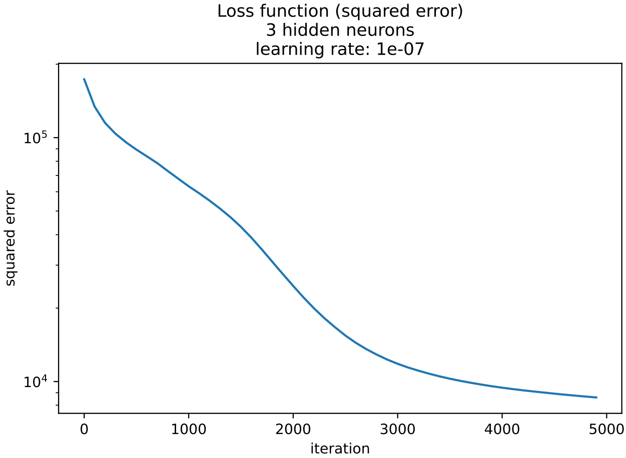
Exercice 1 :

- **cd neural_networks/with_numpy/**. Some toy data are generated in **create_data.py** by a function, + a noise.
- You can tune the standard deviation of the noise
- Use **main_train_neural_network.py** to predict these data by empirical risk minimization.
- You will need to experiment with the **hyperparameters**.
- Make the network find a bad local minimum, make it explode (overflow), observe the network.

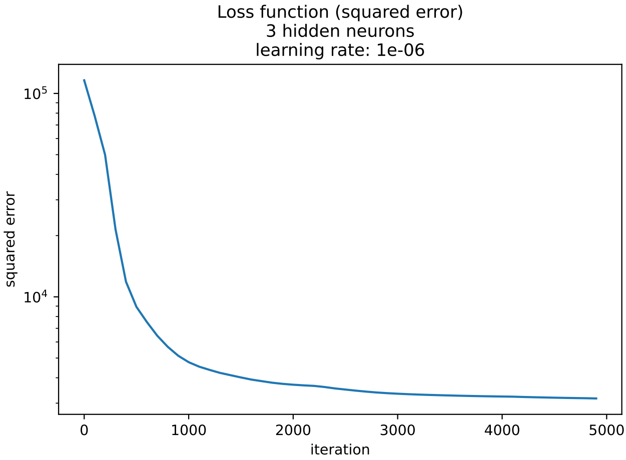
Learning curve



Learning curve



Learning curve

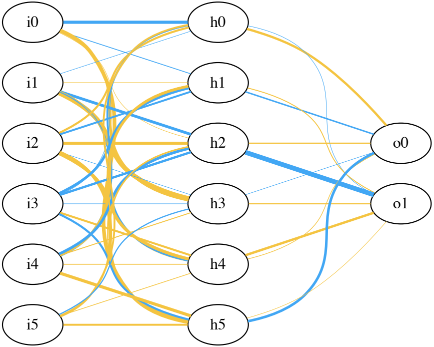


Structured data

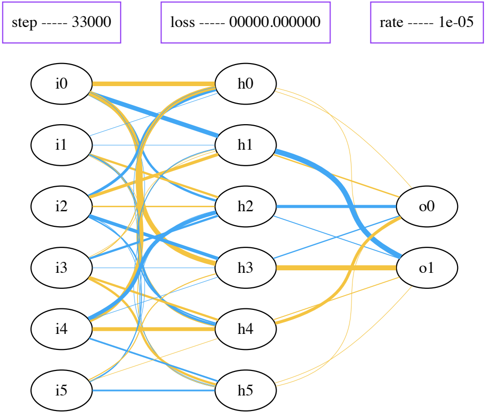
step ----- 00000

loss ----- 135209.182600

rate ----- 1e-05



Structured data



step ----- 33000

loss ----- 00000.000000

rate ----- 1e-05

Plotting

Exercice 2 : **Observation of the network (optional)**

- Uncomment the lines calling **plot_net** so plot the evolution of the network
- You might need to use a smaller network otherwise it will be too long.

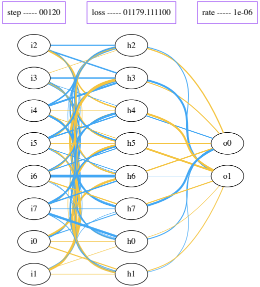
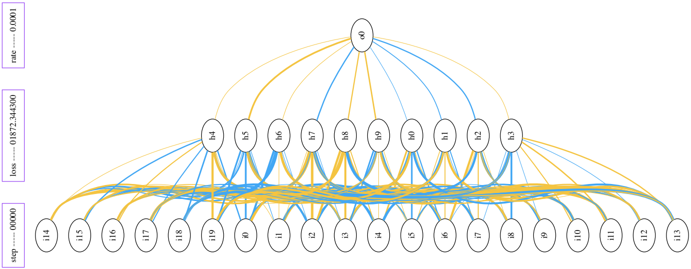
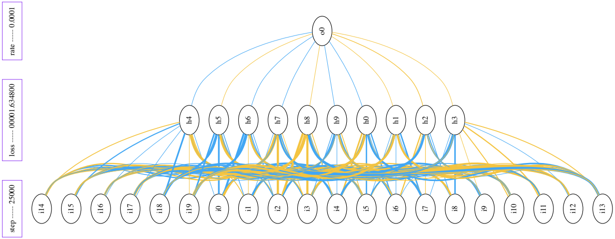


Figure:

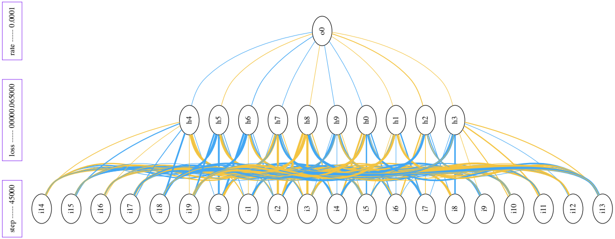
Initial network



After 25000 steps



After 45000 steps



With libs

- We did things manually with numpy but when the networks are large or when we need to automate the seach for good parameters, it is more convenient to work with libraries such as :
  - pytorch
  - tensorflow
  - keras
  - theano

- There are many specific architectures and methods for neural networks:
  - in the number of neurons per layer
  - type of data processed
  - relationship between weights (shared weights)
  - number of hidden layers
  - recurrent neural networks RNN
  - convolutional neural networks CNN
  - graph convolutional networks GCN

Gradient descent

- Stochastic Gradient Descent (SGD)
- Mini-batch learning
- Batch learning

Other cost functions

- Until now we used the squared error cost function
- The slowdown problem
- The Cross entropy is another possible cost function used for classification

MNIST

- **cd ../mnist**
- We will train a neural network to predicts digits in the MNIST database.
- With keras and tensorflow, we will achieve an accuracy of more than 97% in a few minutes for the classification.

Inputs

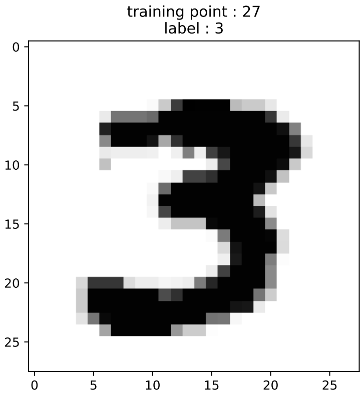


Figure: Datapoint 27

Inputs

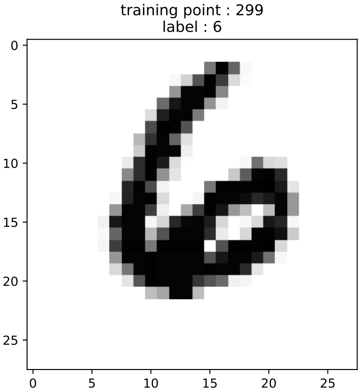
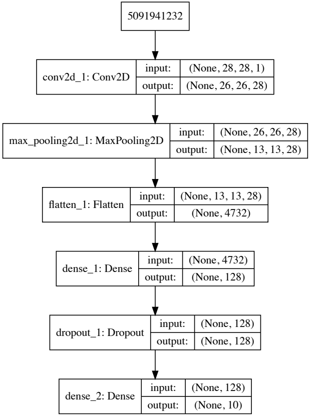


Figure: Datapoint 299

Inputs



5091941232

conv2d_1: Conv2D | input: (None, 28, 28, 1) | output: (None, 26, 26, 28)

max_pooling2d_1: MaxPooling2D | input: (None, 26, 26, 28) | output: (None, 13, 13, 28)

flatten_1: Flatten | input: (None, 13, 13, 28) | output: (None, 4732)

dense_1: Dense | input: (None, 4732) | output: (None, 128)

dropout_1: Dropout | input: (None, 128) | output: (None, 128)

dense_2: Dense | input: (None, 128) | output: (None, 10)

Figure: Network shape

Inputs

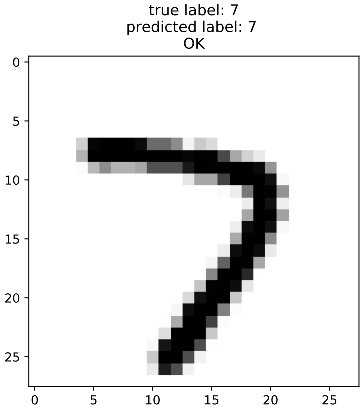


Figure: Predicion for point 17

Inputs

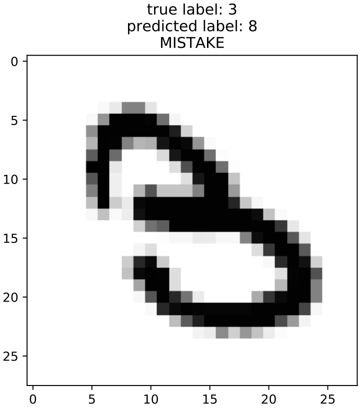
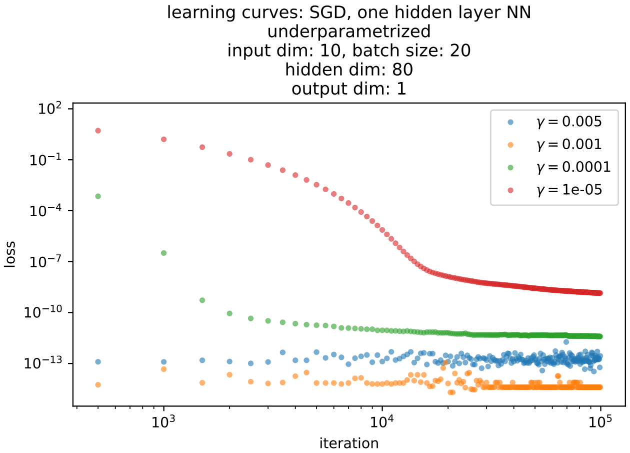
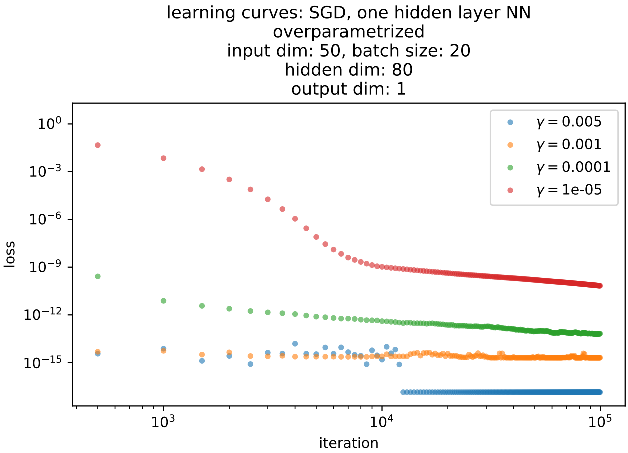


Figure: Prediction for point 18

Underparametrization and overparametrization

Neural network have some very specific behaviors in some contexts.
https://francisbach.com/rethinking-sgd-noise/


learning curves: SGD, one hidden layer NN
underparametrized
input dim: 10, batch size: 20
hidden dim: 80
output dim: 1
10^{2}
10^{-1}
10^{-4}
10^{-7}
10^{-10}
10^{-13}
loss
10^{3}
10^{4}
10^{5}
iteration
$\gamma = 0.005$
$\gamma = 0.001$
$\gamma = 0.0001$
$\gamma =$ 1e-05


learning curves: SGD, one hidden layer NN
overparametrized
input dim: 50, batch size: 20
hidden dim: 80
output dim: 1
$\gamma = 0.005$
$\gamma = 0.001$
$\gamma = 0.0001$
$\gamma =$ 1e-05
10^{0}
10^{-3}
10^{-6}
10^{-9}
10^{-12}
10^{-15}
loss
10^{3}
10^{4}
10^{5}
iteration

Double-descent phenomenon and overfitting

If optimized correctly, neural networks do not overfit easily ! This seems to be in contradiction with the classical machine learning theory.

`https://arxiv.org/pdf/1812.11118.pdf`