Machine Learning Project

Emanuele Cosenza
e.cosenza3@studenti.unipi.it

Riccardo Massidda
r.massidda@studenti.unipi.it

ML course, 2019/2020.
January 29, 2020
Type A project.

Abstract

Design and Python implementation of a multilayer perceptron
trained by means of a minibatch backpropagation algorithm, exploit-
ing various different techniques such as momentum, L2 regularization,
decaying learning rate and early stopping. The network validation is
done through a k-fold cross validation procedure.

Introduction

The presence of different techniques to improve the performances of an
artificial neural network requires the use of formal methods to validate their
effectiveness. Implementing the network and the validation methods from
the ground up has led to the execution of different experiments to motivate
the design choices.

The proposed solution for the competition over the `ML-CUP19` dataset is a
multilayer perceptron designed to be user configurable as much as possible,
allowing a big variety of combinations to be tested independently. The
learning algorithm is based on the backpropagation algorithm[1]. Variations
have been introduced in the update rule to achieve regularization or to
improve the overall performances.

The network also offers the possibility of early stopping since it is a recognized
regularization technique and, furthermore, it reduces the computational time
by not learning for more epochs than required.

Method

For the implementation, Python has been chosen because of its expressiveness and the efficiency of its numerical libraries. In particular, the implementation is based on NumPy[2], which has been used to efficiently manipulate data in form of vectors and matrices. Vectorization has been exploited to speed up the learning process.

Network

The `Network` class represents a neural network. By using its constructor it is possible to set all the required hyperparameters for the techniques that are later described. The class offers methods to learn from a set of examples via backpropagation and to predict outcomes for new patterns in forward mode.

The initialization of the weights in each layer of the network is done by extracting values from a standard normal distribution with variance $\sigma = \frac{2}{n_i + n_o}$, where n_i stands for the number of inputs in the considered layer and n_o for the number of outputs. This has been proven to be a sound choice[3] in various use cases.

Different activation functions can be chosen for each layer of the neural network. The possible choices are: $tanh$, the standard logistic function, $ReLU$ and the identity function, thought to be used only in the output layer for regression tasks.

The implemented backpropagation algorithm analyzes patterns by aggregating them using the minibatch technique. The minibatch size is a tunable hyperparameter with possible values between 1 (online training) and the size of the training set (batch training). In the gradient descent algorithm, MSE is always used as the cost function. To speedup the computation, the update rule also considers momentum information, achieving convergence with a smaller number of epochs. Standard L2 regularization has also been implemented to avoid the overfitting of the training data.

The combination of some hyperparameters could lead to numerical errors due to a gradient explosion phenomenon[4]. This problem is dealt with by normalizing the gradient if it surpasses a certain threshold.

The learning process can be terminated with different stopping criteria, as seen in figure 1. The meaning of the different scenarios is the following:

a. A fixed number of epochs can be provided as an hyperparameter,

leading the network to be trained for no more than the given value.

b. The training process is executed until the loss on the training set reaches a certain provided value.

c. If the loss on the training set does not improve for a fixed number of consecutive epochs, the learning process is stopped.

d. An early stopping mechanism is implemented by checking if the loss on a given validation set does not improve for a fixed number of consecutive epochs. This solution also leads to an implicit regularization of the model, avoiding the overfitting of the dataset[5].

All of this techniques are bounded by a tunable maximum number of epochs, this is needed to avoid situations where there are no assurances about the effectiveness of the stopping criterion like in the case **b**.
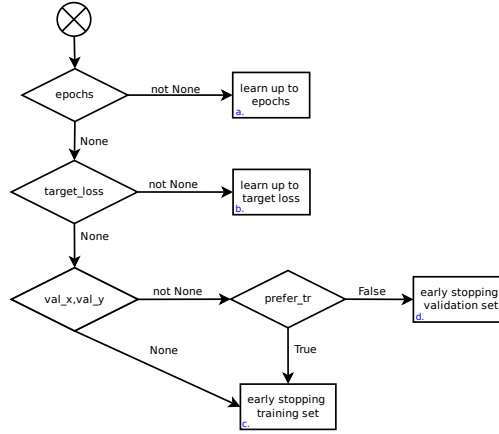

Figure 1: Flow chart for the stopping conditions, the conditions are expressed using the namespace of the implementation.

Validation

The lack of a reliable external test set led to the development of a strategy to assess the performances of the model by using an internal one. Because of the explicit requirement to plot the learning curve of the selected final model against both the training and the test set, double cross validation has been discarded for the assessment, since it would have produced only a scalar value representing the risk of the considered family of models. Given this constraint in the validation procedure, the dataset is partitioned in development set and test set by random sampling without replacement in proportion 80/20%.

The development set is then used for model selection purposes through a cross validation procedure, while the test set is used to assess the selected final model. All the validation methods hereby described are implemented in the `validation.py` script.

The model selection follows a grid search approach implemented in the class `Grid`, capable to perform the Cartesian product over the set of relevant values for each hyperparameter, returning an iterable over all the sound combinations. The grid search is used for model selection, executing in parallel the k-fold cross validation algorithm for each generated assignment.

Given the results of all the cross-validation executions, the model with the minimum average error on the validation sets is chosen as the winner of the process. If early stopping was used inside the cross-validation procedure, the average number of epochs reached or the average error on the training set is returned to allow the final model to be retrained on the development set without a validation set.

Once completed the model selection process, by using the internal test partition extracted from the dataset, it is then possible to assess the final model and obtain the loss information needed to plot the learning curve of the model.

The mechanism hereby described is used in the script `ml-cup.py` to automatically perform model selection and assessment, in the same script, plots and results for the blind competition are produced.

Experiments

MONK's dataset

The results illustrated in table 1 are obtained by averaging eight independent runs for each task as implemented in `monks.py` script. A single layer of four units with *tanh* as the activation function has been used for all the experiments. The output layer is composed by a single unit with a standard logistic activation function. The classification value is obtained by rounding the output up to the nearest integer. The networks have been trained for 2000 epochs by using a minibatch of 32 examples. No further techniques are used in the experiments in Figure 2, Tikhonov regularization is used for the Monks-3 regularized experiment in Figure 3.

4

Table 1: (Experimental results over the MONK's datasets)

| Task	Model	MSE (TR/TS)	Accuracy (TR/TS) (%)
monks-1	$\eta = 0.5$	0.0005/0.0019	100.0%/99.91%
monks-2	$\eta = 0.5$	0.0003/0.0007	100.0%/100.0%
monks-3	$\eta = 0.5$	0.0091/0.0416	99.18%/94.50%
monks-3-reg	$\eta = 0.5, \lambda = 0.01$	0.1160/0.1075	93.44%/97.22%

Cup Results

Screening

A set of preliminary trials, some of which have been automated by the `screening.py` script, have been executed to identify sound ranges for the hyperparameters used for the grid search in the model selection phase. Some of the plots that we observed and discussed in the screening phase are resumed in the appendix.

The α value used to tune the momentum effect is more effective when near its maximum allowed value, 1. While this can generate oscillations in the learning curve, we have observed that it actually smoothens an otherwise noisy curve when using a big learning rate.
Different η values for the fixed learning rate have been observed. We noticed that there is an obvious strict correlation between the η value and the size of the minibatch used for learning. In order to study various configurations we decided to fix the minibatch size and to consequently vary η to improve the performances of the neural network, assuming a practical standard for the minibatch size as 32.

To find reasonable bounds for the decaying learning rate we took advice from literature[6]. We have been able to confirm that good values for τ, the number of iterations required before fixing the learning rate, are in the order of a few hundreds. We then found a good ratio between the initial η_0 and the final η_τ learning rate.

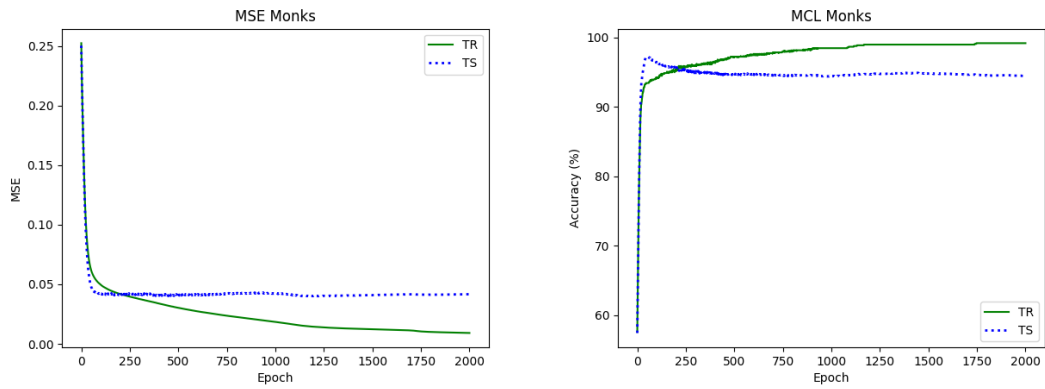The parameter λ used for L2 regularization has been considered effective when set to small values. Considering the fact that early stopping has been used as a stopping criterion in the model selection procedure, and the fact that early stopping is itself a regularization technique, we also allowed a zero value for λ (no L2 regularization).
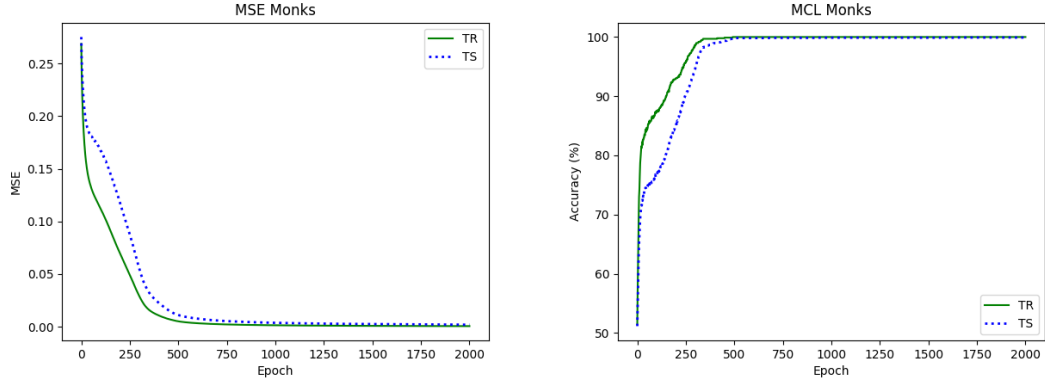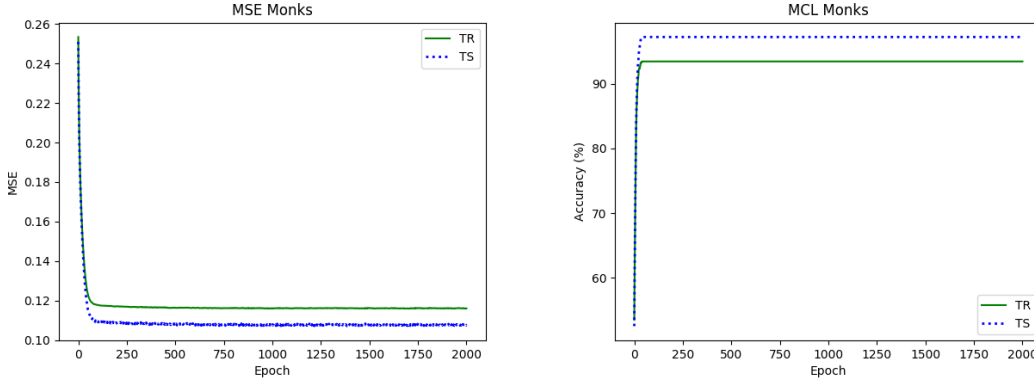
(a) Monks-1
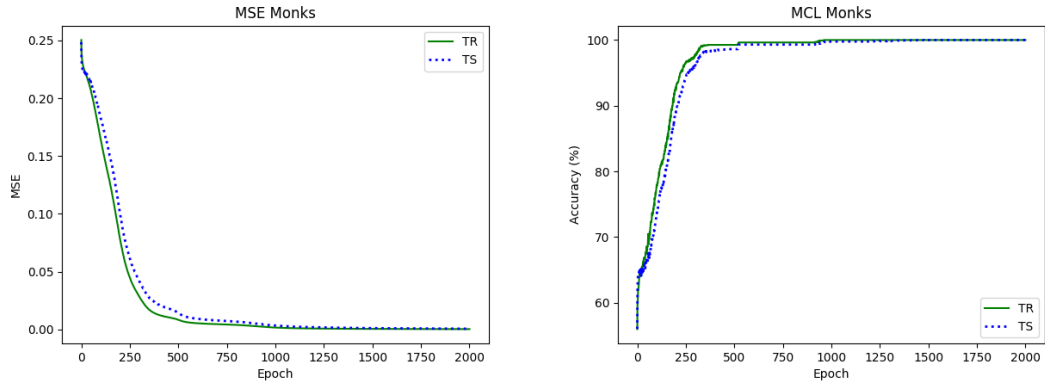
(b) Monks-2

(c) Monks-3

Figure 2: Monks benchmark

(a) Monks-3

Figure 3: Monks-3 with regularization

Regarding the early stopping, we have found an interesting trade-off for the patience parameter in a few hundreds of epochs. This choice was derived from the observation of the relationship between patience and number of epochs and the one between patience and error on the experiment validation set. In particular, we noticed that the number of epochs quickly grows linearly with the patience, while the error on the VL decreases more than linearly with respect to the patience.

In order to obtain a grid of reasonable size, we narrowed the search for the hidden activation function to *tanh* and ReLU. We chose to discard the standard logistic function, as it seemed to always converge slower than *tanh* and ReLU. Since the `ML-CUP19` dataset offers a regression task, the identity function was considered as the only activation function in the output layer.

Given the relative simplicity of the task at hand, we chose to consider only networks composed by one or two layers. To establish reasonable values for the number of units, we experimented with networks having a progressively higher number of units. In accordance to ML theory, we found the more complex networks to be more flexible than the simpler ones. At the end of this phase, we finally considered three different topologies: [20, 32, 2], [20, 64, 2] and [20, 32, 32, 2].

These considerations lead to the final grid chosen for the model selection that we report in table 2. The grid describes 108 possible combinations of hyperparameters, of which 72 with fixed learning rate and 36 with decaying

learning rate.

Table 2: (Range of hyperparameters used in the grid search)

	Hyperparameter	Values
	topology	[20,32,2], [20,64,2], [20,32,32,2]
f	f_hidden	tanh, ReLU
η	eta	$5 * 10^{-2}$, $1 * 10^{-2}$
λ	weight_decay	$1 * 10^{-4}$, $5 * 10^{-5}$, 0
α	momentum	0.99, 0.999
	minibatch	32
	patience	100
τ	tau	200
η_0	eta_zero	0.1
η_τ	eta	0.01

Final model: selection, assessment and discussion

After the screening phase and the grid building procedure, we proceeded to select a model and assess it through a final experiment. As any other experiment we did, this was executed on a machine with an 8 core CPU (Intel i7-8565U (8) @ 4.600GHz) and 8 Gb of RAM, granting a maximum parallelism degree of 8 in the cross validation.

Following the formal procedure described in the validation section, a model has been chosen by means of a grid search using a 10-fold cross validation schema. The entire model selection took 4 hours and 40 minutes to finish. The average training time for a single model was 2 minutes. We report the grid search results for the most relevant models in Table 3.

Table 3: (Results for the grid search. Only the most relevant models are reported, each with its hyperparameters, training MEE and validation MEE. Training MEE and validation MEE are computed as averages on the 10 folds.)

	topology	f	η	α	λ	TR MEE	VL MEE
Winner	[20, 32, 32, 2]	$tanh$	$5 * 10^{-2}$	0.99	$1 * 10^{-4}$	0.834	1.008
2nd	[20, 32, 32, 2]	$tanh$	$5 * 10^{-2}$	0.99	$5 * 10^{-5}$	0.836	1.013
3rd	[20, 32, 32, 2]	$tanh$	$1 * 10^{-2}$	0.99	$5 * 10^{-5}$	0.924	1.024
11th	[20, 32, 32, 2]	$relu$	$5 * 10^{-2}$	0.99	$5 * 10^{-4}$	0.914	1.061

	topology	f	η	α	λ	TR MEE	VL MEE
13th	[20, 32, 2]	$tanh$	$5*10^{-2}$	0.99	$5*10^{-4}$	0.952	1.063

The topology [20, 32, 32, 2] was by far the most successful (the best network with a different topology finished in 13th position). Similarly, *tanh* was by far the best hidden activation function (the best ReLU network placed 11th). Also, $\alpha = 0.99$ happened to be a good value for the momentum parameter, while the η parameter did not seem to be decisive with respect to generalization capacity.

Given the selected final model, we retrained it on the development set (80% of the data) and we then assessed it on the previously extracted internal test set (20% of the data). This last retraining was stopped as soon as the training loss reached the average training loss reached by the best model in the model selection phase. The plot relative to the final retraining, which contains TR and TS learning curves, is reported in Figure 4, while the MEE results on the overall validation process are in table 4.

Table 4: (MEE for the chosen model over the partitions of the dataset)

TR	VL	TS
0.834	1.008	1.019

The plot shows again that the selected model converges well on the training set, proving that the backpropagation algorithm has been implemented correctly. Also, the validation error goes down together with the training error, showing that the network is able to generalize on unseen data.

The final trained model has been used to compute the final outputs for the blind test set. By representing them as points in the output space plot shown in figure 5 in the appendix, we found that the outputs of our final model lied near the curve identified by the training outputs, showing that the model approximated well the unknown function.

Conclusions

In this report we presented a NumPy implementation of a multilayer perceptron equipped with momentum, L2 regularization and various configurable
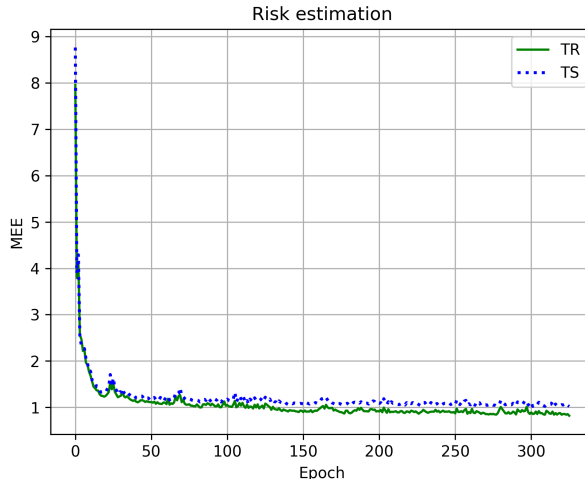
Figure 4: Learning curves (TR and TS) for the final retraining.

stopping criteria, including early stopping. We reported the main design choices, describing the network structure and the main techniques involved in it.

We then described the general procedures used to validate the network: for model selection we implemented a grid search procedure over a k-fold cross validation, while for model assessment we used a simple hold out division (80%/20%). Parallelization has been exploited to reduce the computation time.

We showed the proper functioning of the learning algorithm on the MONKS dataset, proving that the network can learn and generalize well in a classification task.

Next, we focused on the proposed `ML-CUP19` competition. In the screening phase, we analysed the behaviour of the model under different circumstances on the regression dataset, doing various experiments to find good ranges for the tunable hyperparameters. With a model selection procedure, we found out that two-layered networks perform better than networks with a single layer. Similarly, *tanh* happened to be the best choice with respect to the hidden activation function. Finally, we trained a final network assessing it on an internal test set, giving an estimate of the network performance on the blind test set.

10

References

1. Rumelhart, D. E. & McClelland, J. L. *Parallel distributed processing: Explorations in the microstructure of cognition.* (MIT Press, 1986).

2. Oliphant, T. E. *Guide to NumPy.* (Continuum Press, 2015).

3. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. 8.

4. Pascanu, R., Mikolov, T. & Bengio, Y. On the difficulty of training recurrent neural networks. (2012).

5. Prechelt, L. Early stopping but when? 15.

6. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning.* (The MIT Press, 2016).

Appendix

Output space
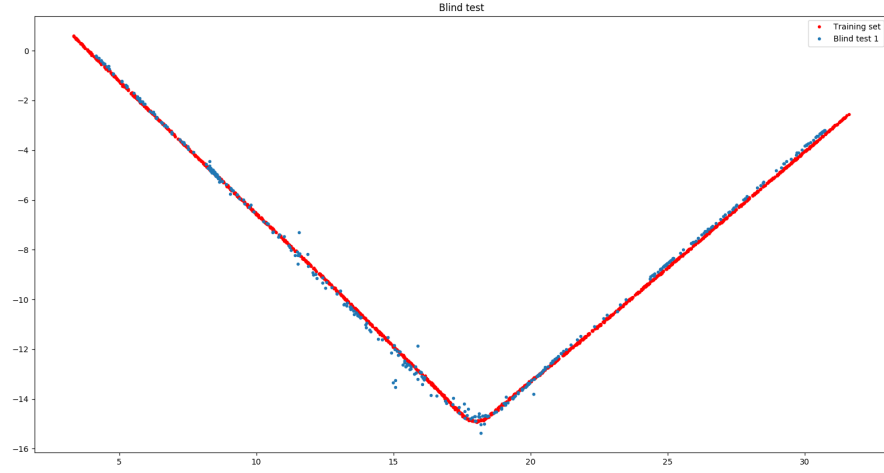
Figure 5: Plot in the output space of the results on the blind test set.

Screening

"Fixed learning rate" consists of 8 models

Model	epochs	η	topology	MSE
0	200	1	[20 15 2]	15.295817557113532
1	200	0.5	[20 15 2]	3.112441820482686
2	200	0.1	[20 15 2]	2.2542645414631917
3	200	0.05	[20 15 2]	2.521080318408892
4	200	0.01	[20 15 2]	3.2346240986146078
5	200	0.001	[20 15 2]	10.618540535255717
6	200	0.0001	[20 15 2]	198.46763156549304
7	200	1e-05	[20 15 2]	435.654983785013

"Decay learning rate" consists of 4 models

Model	epochs	η	η_0	τ	topology	MSE
0	200	0.005	0.5	100	[20 15 2]	3.5727348531513345
1	200	0.005	0.5	200	[20 15 2]	3.222256412989206
2	200	0.001	0.1	100	[20 15 2]	4.136002939917029
3	200	0.001	0.1	200	[20 15 2]	3.3292547553770895

"Oscillating decay learning rate" consists of 4 models

Model	epochs	η	η_0	τ	topology	MSE
0	200	0.05	5	100	[20 15 2]	1.7643450551336316e+36
1	200	0.05	5	200	[20 15 2]	6.186715674416807e+90
2	200	0.01	1	100	[20 15 2]	5.068823772090453
3	200	0.01	1	200	[20 15 2]	4.909584508252798

"Minibatch-bigstep" consists of 4 models

Model	epochs	η	minibatch	topology	MSE
0	200	1	1	[20 15 2]	731.6221925869986
1	200	1	4	[20 15 2]	306.35193748538717
2	200	0.5	1	[20 15 2]	365.63750367443646
3	200	0.5	4	[20 15 2]	180.22096643788132

"Tikhonov regularization (L2)" consists of 5 models

Model	epochs	topology	λ	MSE
0	200	[20 15 2]	0	3.256676585056649
1	200	[20 15 2]	0.1	73.46138718129757
2	200	[20 15 2]	0.01	14.338679697054966
3	200	[20 15 2]	0.001	3.9800649118526756
4	200	[20 15 2]	0.0001	3.1959960722144527

"Momentum" consists of 5 models

Model	epochs	α	topology	MSE
0	200	0	[20 15 2]	3.2583345038923106
1	200	0.5	[20 15 2]	3.172221208044486
2	200	0.9	[20 15 2]	3.2819828404574456
3	200	0.99	[20 15 2]	3.4393323598920507
4	200	0.999	[20 15 2]	31.813636105146752

"Gradient clipping" consists of 4 models

Model	epochs	max_norm	topology	MSE
0	200	1	[20 15 2]	3.2479584177805774
1	200	2	[20 15 2]	3.269801785468375
2	200	10	[20 15 2]	3.357096724331396
3	200	100	[20 15 2]	1.8067783005117501

"Activation functions" consists of 3 models

Model	epochs	f_h	topology	MSE
0	200	'tanh'	[20 15 2]	3.3506727028754204
1	200	'sigmoid'	[20 15 2]	4.30754127358549
2	200	'relu'	[20 15 2]	3.4702944013758112

"Assess different values of patience," consists of 10 models

Model	patience	prefer_tr	topology	epochs	MSE
0	1	False	[20 15 2]	39	13.649828858564003
1	2	False	[20 15 2]	71	9.358610710765161
2	4	False	[20 15 2]	154	3.7510464105064387
3	8	False	[20 15 2]	207	3.564903922675451
4	16	False	[20 15 2]	309	3.136955126870644
5	32	False	[20 15 2]	236	3.136569191859081
6	64	False	[20 15 2]	248	3.2515729733888405
7	128	False	[20 15 2]	373	3.146595153279893
8	256	False	[20 15 2]	563	3.085762997256863
9	512	False	[20 15 2]	623	2.955754499161358

"Topology" consists of 5 models

Model	epochs	η	topology	MSE
0	200	0.05	[20 5 2]	4.202882018807765
1	200	0.05	[20 32 2]	2.2692850005667955
2	200	0.05	[20 64 2]	2.2545935152610777
3	200	0.05	[20 32 32 2]	1.9399796521701196
4	200	0.05	[20 32 32 32 2]	2.0031332556940007