

# MODEL ORDER REDUCTION OF RAREFIED GASES USING NEURONAL NETWORKS

## Bachelorarbeit

zur Erlangung des akademischen Grades  
Bachelor of Science (B. Sc.)  
im Fach Physikalische Ingenieurwissenschaften



Technische Universität Berlin  
Fakultät Verkehrs- und Maschinensysteme V  
Institut für Numerische Fluidodynamik

eingereicht von: *Zachary Schellin*  
geboren am: *11.02.1991, Berlin*

Gutachter: *Prof. Dr. Julius Reiss*  
*Dr. Mathias Lemke*

eingereicht am: *19. Februar 2021*

## **Abstract**

Abstract here

## **Zusammenfassung**

german abstrac here

# Contents

---

<b>Contents</b>	<b>i</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Objective of this thesis . . . . .	1
1.2 Thesis outline . . . . .	1
1.3 State of the art . . . . .	2
<b>2 The BGK Model</b>	<b>3</b>
2.1 Discretization, conservative properties and equilibrium . . . . .	3
2.2 Sod's schock tube as a test case for the BGK model . . . . .	4
<b>3 Dimensionality reduction algorithms</b>	<b>6</b>
3.1 Proper orthogonal decomposition (POD) . . . . .	6
3.2 Autoencoders . . . . .	6
3.2.1 Autoencoders . . . . .	9
<b>4 Model Order Reduction</b>	<b>10</b>
4.1 Offline Phase . . . . .	11
4.1.1 Full order BGK-model . . . . .	11
4.1.2 Constructing the mapping . . . . .	12
4.1.3 Online Phase . . . . .	13
4.1.4 Reduced Order Model . . . . .	13
<b>5 Results</b>	<b>15</b>
5.0.1 Discussion and Outlook . . . . .	20
<b>A Hyperparameters for the Fully Connected Autoencoder</b>	<b>26</b>
<b>B Hyperparameters for the Convolutional Autoencoder</b>	<b>41</b>
B.0.1 Appendix B . . . . .	41
<b>Bibliography</b>	<b>44</b>

# **1 Introduction**

---

The Bhatnagar-Gross-Krook equation (BGK) is a kinetic collision model of ionized and neutral gases valid for rarefied as well as other pressure regimes [1]. Generating data of such a flow field is essential for various industry and scientific applications[REF]. With the intention to reduce time and cost during the data generating process, experiments were substituted with computational fluid dynamics (CFD) computations. Consequently reduced-order models (ROMs) coupled to aforementioned computations were introduced to further the reduction of time and cost. The thriving field of artificial intelligence operates in model order reduction for data visualization/analysis since the 80's (Quelle?) and has now surfaced in fluid mechanics. This thesis will cover the use of artificial intelligence for model order reduction in fluid mechanics.

## **1.1 Objective of this thesis**

Due to the non-linearity of transport problems in particular shock fronts, the construction of a robust ROM for those cases poses several challenges.

## **1.2 Thesis outline**

1. What is the BGK Model
2. What is the SOD and BGK in SOD
3. What is deep learning what are autoencoders
4. hyperparameters for autoencoders
5. What is a reduced order model
6. offline phase
7. my data in 1d bgk in sod shock tube FOM Data
8. what is pod and Rb by POD
9. RB by autoencoder -online phase
10. what is my ROM
11. results - results by ROM for FC and for CONV

12. variation of intrinsic variables
13. Comparison to POD intrinsic variables number and quality

### 1.3 State of the art

State of the art model reduction of dynamical systems can be done via proper orthogonal decomposition (POD) which is an algorithm feeding on the idea of singular value decomposition (SVD)[2][3]. POD captures a low-rank representation on a linear manifold. So called POD modes, derived from SVD, describe the principle components of a problem which can be coupled within a Galerkin framework to produce an approximation of a lower dimension  $r$ .

$$f(x) \approx \tilde{f}(x) \quad \text{with } \tilde{f} = \sum_{k=1}^r a_k \psi_k(x) \quad \text{where } \psi_k \text{ are orthonormal functions.} \quad (1.1)$$

Bernard et al. use POD-Galerkin with an additional population of their snapshot database via optimal transport for the proposed BGK equation, bisecting computational run time (cost) in conjunction with an approximation error of  $\sim 1\%$  in [4]. Artificial intelligence in the form of autoencoders replacing the POD within a Galerkin framework is evaluated against the POD performance by Kookjin et al. for advection-dominated problems[5] resulting in sub 0.1% errors. An additional time inter- and extrapolation is evaluated. Using machine learning/ deep learning for reduced order modeling in CFD is a novel approach although "the idea of autoencoders has been part of the historical landscape of neural networks for decades"[6, p.493]. Autoencoders, or more precisely learning internal representations by the delta rule (backpropagation) and the use of hidden units in a feed forward neural network architecture, premiered by Rumelhart et al. (1986) [7]. Through so called hierarchical training Ballard et al.(1987) introduce a strategy to train auto associative networks (nowadays referred to as autoencoders), in a reasonable time promoting further development despite computational limitations [8]. The so called bottleneck of autoencoders yields a non-smooth and entangled representation thus being uninterpretable by practitioners[9] leading to developments in this field. Rifai et al. introduce the contractive autoencoder (CAE) for classification tasks (2011), with the aim to extract robust features which are insensitive to input variations orthogonal to the low-dimensional non-linear manifold by adding a penalty on the frobenius norm of the intrinsic variables with respect to the input, surpassing other classification algorithms [9]. Subsequent development emerges with the manifold tangent classifier (MTC) [10]. A local chart for each datapoint is obtained hence characterizing the manifold which in turn improves classification performance. On that basis a generative process for the CAE is developed. Through movements along the manifold with directions defined by the Jacobian of the bottleneck layer with respect to the input  $\vec{x}_m = JJ^T$ , sampling is realized [11].... Proper orthogonal decomposition (POD) and its numerous variants like shifted-POD[?], POD-Galerkin[?], POD+I [?] to name only a few of them, try to solve this problem by.....

## 2 The BGK Model

---

This section covers the kinetic gas model, the BGK model, on which a model order reduction will be performed in the following. In addition the SOD-shocktube, on which the BGK model will be tested is discussed.

### 2.1 Discretization, conservative properties and equilibrium

The BGK model was introduced by, and named after, physicists Prabhu L. Bhatnagar, Eugene P. Gross and Max Krook in 1954 [1]. It is an approximation of the standard Boltzmann transport equation. More precisely the r.h.s of the Boltzmann equation is approximated by the BGK operator [12]

$$\partial_t f + v \partial_x f = \frac{1}{\tau} (M_f - f). \quad (2.1)$$

It consists of the relaxation time  $\tau(x, t)$ , the Maxwellian distribution  $M_f$  and  $f(x, v, t)$  the probability of a gas particle having a microscopic velocity  $v$  in phase space  $(x, v, t)$ . The left side of the BGK model is a transport equation for  $f(x, v, t)$  with transport velocity  $v$ . In the Maxwellian distribution  $n(x, t)$  is the number of particles in space and time,  $R$  is the specific gas constant,  $T^0(x, t)$  is the temperature and  $u(x, t)$  is the macroscopic velocity

$$M_f = \frac{n(x, t)}{\sqrt{2\pi RT^0(x, t)}} \exp\left(-\frac{(v - u(x, t))^2}{2RT^0(x, t)}\right). \quad (2.2)$$

Note that in this thesis the BGK model is discussed in one dimension. Hence the BGK model needs to be evaluated for three independent variables  $x$ ,  $v$  and  $t$  as seen above.

Now what makes the BGK model especially attractive for model order reduction? The fruitfulness of performing a model order reduction on the BGK model becomes clearer when looking at its space and velocity discretization

$$\partial_t f_{j,k} = -(v_k)_1 D_x f|_{j,k}(t) + \frac{1}{\tau} (M_{f,j,k}(t) - f_{j,k}(t)). \quad (2.3)$$

Here a uniform grid is considered with  $x_j = j\Delta x$ ,  $j \in \mathbb{Z}$ ,  $v_k = k\Delta v$ ,  $k \in \mathbb{Z}$  and  $t^n = n\Delta t$ ,  $n \in \mathbb{N}$  on which  $f_{j,k} = f(x_j, v_k, t)$  and  $M_{f,j,k} = M_f(x_j, v_k, t)$  are evaluated at point  $(x_j, v_k)$  in a time instance  $t$ . For brevity  $D_x f|_{j,k}$  is the discrete space derivative at  $(x_j, v_k)$  [12]. Now the partial differential equation (PDE) in eq. (2.1) is broken down into a system of ordinary differential equations (ODE's) in time, for which every ODE is a linear advection equation with constant scalar speed  $v_k$  and a source term.

To continue let's consider  $K$  to be the number of gridpoints in velocity and  $J$  to be the number of grid points in space. Then a total of  $KJ$  first order differential equations need to be evaluated in 1D. Obviously in three dimensions the system of ODE's inflates up to  $K^3N^3$  first order differential equations eq. (2.3). This in turn drives the evaluation of the BGK model at the edge of intractability for dense meshes in 3D and all together stimulate the want for a reduced order model.

In order to evaluate  $M_f$  and therefore eq. (2.3)

the macroscopic velocity  $u(x, t)$  is required.

Other macroscopic quantities of the gas flow, namely the density  $\rho$ , the momentum  $\rho u$  and the energy  $E$  are the expected values or moments of  $f$  in velocity space and can be calculated with

$$\rho(x, t) = m \int f dv, \quad (2.4)$$

$$\rho(x, t)u(x, t) = m \int vf dv, \quad (2.5)$$

$$E(x, t) = m \int \frac{1}{2}v^2 f dv, \quad (2.6)$$

as seen in [12]. Here  $f$  is integrated over velocity space and multiplied by  $m\Phi$  with  $\Phi = [1, v, \frac{1}{2}v^2]$ , the collision invariants and  $m$  the mass of the particles.

Displayed in fig. 2.1 is a demonstrative example of how the distribution function  $f(v)$  gives the values for the macroscopic quantities. The

distribution is centered around the macroscopic velocity  $u$ , the mean velocity of the distribution  $f$  is the temperature  $T$ , integrating  $f(v)$  over velocity space one obtains the density  $\rho$ .

Evidently the system in eq. (2.1) is in equilibrium when  $f = M_f$ . Now multiplying the equilibrium solution of eq. (2.1) by  $m\Phi(v)$  and integrating in velocity space, one finds the Euler system of classical gas dynamics using the equation of state of the gas [12]

$$\partial_t \rho + \partial_x(\rho u) = 0 \quad (2.7)$$

$$\partial_t(\rho u) + \partial_x(\rho u^2 + p) = 0 \quad (2.8)$$

$$\partial_t E + \partial_x(u(E + p)) = 0. \quad (2.9)$$

Note that the Boltzmann transport equation the conservative properties of the microscopic collisions lead to a global conservation of mass, momentum and energy. Hence the conservation of mass, momentum and energy is found in the BGK model as well [12].

To continue ..... rarefaction KN SOD

## 2.2 Sod's shock tube as a test case for the BGK model

- intrinsic code variables for hydro an rare add pls

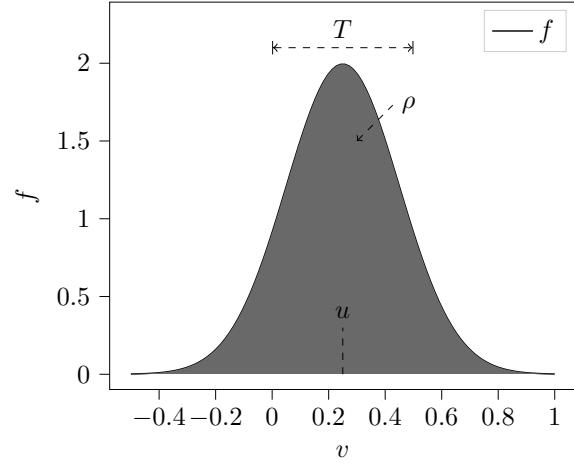


Figure 2.1: Illustration of the linkage between the macroscopic quantities of the gas flow and  $f$  the distribution function  $f$ .

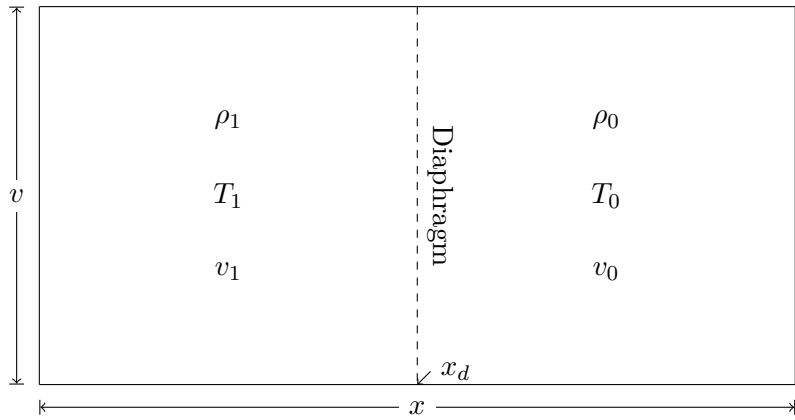


Figure 2.2: Problem setup for the Sod shock tube in 1D. A diaphragm is positioned at  $x_d$ , separating the whole domain in two areas. Shown are the starting conditions for  $\rho, v$  and  $T$  indexed with one. Zero indexing refers to the absence of matter.

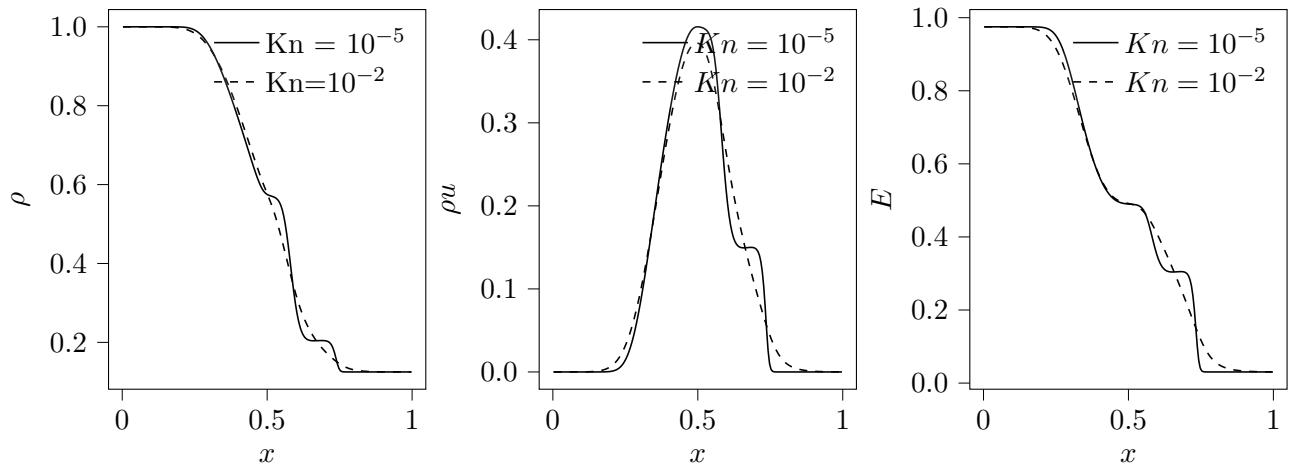


Figure 2.3: Macroscopic quantities  $\rho$ ,  $\rho u$  and  $E$  in the SOD shock tube at the last time stamp. The quantities are displayed for the hydrodynamic regime in - - lines and for the rarefied regime in - lines.

## **3 Dimensionality reduction algorithms**

---

### **3.1 Proper orthogonal decomposition (POD)**

The singular value decomposition of the input  $X$  [REF to Section 1] gives the optimal low-rank approximation  $\tilde{X}$  of  $X$  ??[Eckard-Young].

$$\underset{\tilde{X}, s.t. rank(\tilde{X})=r}{\operatorname{argmin}} \quad ||X - \tilde{X}||_F = \tilde{U}\tilde{\Sigma}\tilde{V}^* \quad (3.1)$$

### **3.2 Autoencoders**

In this section deep learning with the focus on fully connected autoencoders and convolutional autoencoders will be introduced. Starting off with addressing important terminology whilst presenting the concept of autoencoders and continuing with the introduction of the fully connected and convolutional forwardpass , this section closes with ADAM [?] an update to the backpropagation algorithm as well as important training methods.

The term deep learning situated around the much broader field of artificial intelligence stems from the use of deep feed forward networks also called multi layer perceptrons (MLP) or feed forward neural networks. Deep recurrent neural networks (RNN) are also used in this field but won't be covered in this thesis. In contrast to RNNs, information flows forward through these networks, which explains the name feed forward. In the following I will use the abbreviation MLP when talking about the aforementioned algorithm. Network refers to the typical composition of many different functions.

The task of any MLP is to approximate a function  $f^*(\mathbf{x}; \Theta) \approx f(\mathbf{x})$  through learning the values of the parameters  $\Theta$ . As mentioned before  $f^*$  is a composition of functions eg. eq. (3.2).

$$f^*(\mathbf{x}; \Theta) = f^3(f^2(f^1(\mathbf{x}; \Theta^1), \Theta^2), \Theta^3) \quad (3.2)$$

In eq. (3.2) each function  $f^i$  is called layer. In this example  $f^1$  is called the input layer,  $f^3$  the output layer and  $f^2$  a hidden layer. Hence a layer is a vector-to-vector function. In this context a unit describes the corresponding vector-to-scalar functions of one layer. The width of a layer is referred to as the dimension of the vector valued input. The depth of the network describes the number of composed functions. In autoencoders the dimensions of input and output layer are identical.

Autoencoders are a special kind of neural network, that have a central hidden layer, that outputs a code  $c$  which should contain useful features of the input  $x$  while usually being of a lower dimension. Hereafter the code will be addressed as intrinsic variables highlighting the property of containing useful features of the input  $x$ . Autoencoders can be split in two parts, the encoder  $h(x) = c$

which compresses the input and outputs the intrinsic variables  $c$  and the decoder  $g(h) = \hat{x}$  which reconstructs the input from the intrinsic variables to output  $\hat{x}$ . The goal of autoencoders conflicts with the training objective. The former is to produce a code that describes the intrinsic features of the input, while the latter is to minimize the difference between  $x$  and  $\hat{x}$ . Therefore autoencoders need to be restrained from learning the identity function perfectly which in turn should drive the model to choose which instance to copy.

Obviously deep learning emphasizes the focus on the depth of a model. This is because linear models with just one layer can only approximate linear functions. Adding a non-linear activation function to the output of the proposed model wouldn't be sufficient in modeling any nonlinear behavior of the function. However the universal approximation theorem [13] states that MLPs with at least one hidden layer and any non-linear activation function can approximate any function given that enough hidden layers can be provided. In conclusion, MLPs are universal approximators [6].

There are several types of layers, that can be used in MLPs. For this thesis it is sufficient to treat so called *fully connected layers* and *convolutional layers*. Fully connected layers are called **Linear**[?] in PyTorch[?] because they compute a linear transformation of the input eq. (3.3), where  $x$  is the input vector,  $A$  is the weight matrix and  $b$  is a bias vector. This is the forward pass of a linear layer. The learnable parameters  $\Theta$  are in this case the values in  $A$  and  $b$ . For a linear layer which takes a vector of size  $i$  as input and outputs a vector of size  $o$  there are  $l = i \times o + o$  learnable parameters.

$$y = xA^T + b \quad (3.3)$$

Continuing with the forward pass in convolutional layers, which are called **Conv2D**[?] in Pytorch. Convolutional layers are usually applied when the input data has a known grid-like topology [6]. While for fully connected layers, the input size is fixed, convolutional layers can be applied to inputs of various sizes. Furthermore, they are sparse by construction and share parameters making them equivariant [6]. Even though the name implies the use of the convolutional operation in eq. (3.4), PyTorch and many other neural network libraries instead use the cross correlation, which is an operation closely related to a convolution[6][?]. The convolution in 3.4 operates on two functions  $x$  and  $w$ , where the latter is a weighting function of the former which makes  $s$  the weighted average of the input  $x$ . In eq. (3.5)  $x$  is represented by  $I(m, n)$ , and  $w$  by  $K(m, n)$  respectively. Note that while eq. (3.4) is scalar valued and eq. (3.5) is two dimensional, only for illustrative reasons.

$$s(t) = (x * w)(t) = \int x(a)w(t - a) da \quad (3.4)$$

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(m, n)K(i - m, j - n) \quad (3.5)$$

One drawback in implementing the discrete convolution eq. (3.5) is that there can be invalid values for  $m$  and  $n$ . This can be solved by exploiting the commutative property of the convolution, which results for the discrete convolution in a flipped kernel relative to the input, eq. (3.6). Without the need of flipping the kernel which is not always possible, i.e. exploiting the commutative property, cross correlation is adopted eq. (3.7).

$$S(i, j) = (K * I)(i, j) = \sum_m \sum_n I(i - m, j - n)K(m, n) \quad (3.6)$$

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i + m, j + n)K(m, n) \quad (3.7)$$

The given equations illustrate the movement of the kernel over a two dimensional input, but leaving out two basic accompanying designs. First is the so called stride of the kernel, which results in a increased down sampling with increased stride size. Considering strides of the kernel along one dimension of the input results in a shrinkage of that dimension by

$$o = \frac{i - k}{s} + 1. \quad (3.8)$$

Here  $o$ ,  $i$ ,  $s$  and  $k$  are the output, input, stride and kernel size of one dimension respectively [14]. Second are the so called channels, which allow convolutional layers to extract a different feature for every channel from the same input. For a two dimensional input like images this could be different features for the same location on the image, like edges and color in the RGB color space [6]. Adding strides and kernel to eq. (3.7) yields

$$S_{i,j,k} = c(K, I, s)_{i,j,k} \sum_{l,m,n} \left[ I_{l,(j-1) \times s+m, (k-1) \times s+n} K_{i,l,m,n} \right]. \quad (3.9)$$

The kernel  $K$  is a four dimensional tensor with  $i$  indexing into the output channels of  $S$ ,  $l$  indexing into the input channels of  $I$ ,  $m$  and  $n$  indexing into the rows and columns. The input  $I$  and output  $S$  are three dimensional tensors with  $j$  and  $k$  indexing into the rows and columns. Note that in PyTorch eq. (3.9) is a so called valid cross correlation (valid convolution) [?] meaning the kernel will only move over input units for which all  $m$  and  $n$  are inside the rows and columns of the input. Zero padding the input can prevent the kernel from omitting corners in that case.

For autoencoders the necessity to perform a transposition of the applied layers arises, which is straight forward for fully connected layers. Convolutional layers with strides greater than unity on the other hand the transposition needs the kernel to be padded with zeros to realize an upsampling of the input data [14]. Note that the padding of the kernel with zeros is only used to illustrate how the upsampling works. In eq. (3.10) taken from [6] multiplications with zero are omitted. The size of one dimension during upsampling can be calculated with the transposition of eq. (3.8).

$$t(K, H, s)_{i,j,k} = \sum_{\substack{l,m \\ s.t. \\ (l-1) \times s+m=j}} \sum_{\substack{n,p \\ s.t. \\ (n-1) \times s+p=k}} \sum_q K_{q,i,m,p} H_{q,l,n} \quad (3.10)$$

Forward-propagation is then referred to as the compositional evaluation of each layer. For the example in eq. (3.2) the outputs of each layer would then be:  $f^1(\mathbf{x}, \Theta^1) = \mathbf{p}$ ,  $f^2(\mathbf{p}, \Theta^2) = \mathbf{q}$  and  $f^3(\mathbf{q}, \Theta^3) = \hat{\mathbf{y}}$ .

Subsequently the cost function  $J(\Theta)$ , which will be discussed later on, can be computed. Afterwards back-propagation returns the gradients w.r.t. the layer parameters  $\Theta$  to finally compute updated parameters  $\Theta$ . The name back-propagation refers to the use of the chain rule of calculus to obtain the gradients of each layer. Assuming again the example in eq. (3.2) back-propagation would be equations eq. (3.13) to eq. (3.11):

$$\frac{\partial J}{\partial \Theta^3} = \frac{\partial J}{\partial \hat{y}_i} \frac{\partial \hat{y}}{\partial \Theta^3} \quad (3.11)$$

$$\frac{\partial J}{\partial \Theta^2} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial q} \frac{\partial q}{\partial \Theta^2} \quad (3.12)$$

$$\frac{\partial J}{\partial \Theta^1} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial p} \frac{\partial p}{\partial q} \frac{\partial q}{\partial \Theta^1} \quad (3.13)$$

While the term back-propagation is solely used for the method to compute the gradients for each layer in a backward fashion, meaning from the last layer to the first, the update of the parameters is done in an optimization step. -L2-error as performance metrics

### 3.2.1 Autoencoders

Autoencoders have many hyperparameters determining their capability for compression and subsequent reconstruction. These parameters include : *depth, width of layers, activation functions, batch-size, learning rate, number of filters, stride width, kernel size*. Their finding is discussed in this section, for both **H** and **R**.

For the fully connected autoencoder the order of determining the hyperparameters is as follows: Depth → Hidden Units → Batch Size → Activation Functions. For the convolutional autoencoder the order of determining the hyperparameters is as follows: Depth → Channels → Batch Size → Activation Functions. Both models are evaluated on the course of the training. Figures showing the training process for all experiments are provided in the appendix. After training the L2-Error?? is evaluated on the unshuffled, complete dataset as described in ??.

Important to mention is that the difficulty of finding the right set of hyperparameters for MLPs led to an intensive search prior to the creation of this contribution. The difficulty lies in the fact that there is little systematic knowledge about how the hyperparameters interact in the model and answer to a variety of input data. Goodfellow et al. point out that with a combination of intuition, certain methods and first of all experience practitioners find hyperparameters that work well [6]. As for this thesis a list of the prior search is provided in appendix B.0.1. This list does not claim to cover the complete search and can by no means give enough insight to enable reproducability by another person.

For this reason an insight into the methods employed to find the set of hyperparameters used in this contribution are described in the following.

The convolutional autoencoder architecture 1.1 is a composition of six convolutional and three fully connected layers, eq. (3.14).

$$y_p = f_C^9(f_C^8(f_C^7(f_C^6(f_F^5(f_F^4(f_C^3(f_C^2(f_C^1(y_0)))))))))) \quad (3.14)$$

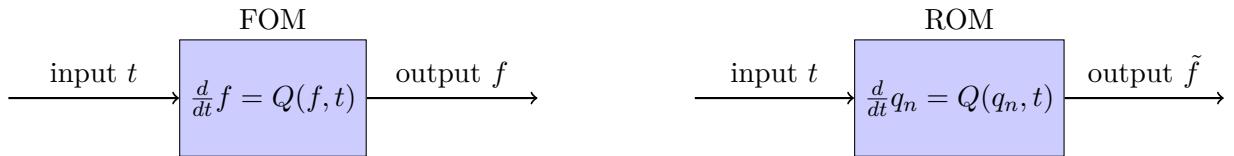
## 4 Model Order Reduction

---

In this chapter model order reduction (MOR) of the BGK-model in the Sod shock tube will be introduced for which POD and in particular autoencoders are adopted to obtain a reduced basis (RB).

Model order reduction is a technique used for reducing the computational cost when evaluating PDE's [4][5][15]. To achieve this, the solution to the PDE is being approximated by reducing one or more of it's dimensions. The reduction is performed by a mapping onto a low dimensional manifold. In our case the solution to the BGK-model is  $f(x, v, t) \in \mathbb{R}^3$ . Now we could reduce i.e.  $v$  to  $n$ , with  $o$  beeing the number of elements in  $v$  and  $p$  in  $n$ . With  $p \ll o$  we obtain a reduced order model (ROM)  $q(x, n, t)$  of significantly lower dimension. In particular  $n$  is called the reduced basis or the intrinsic variables. In chapter 2 we saw that the BGK-model is a PDE which through discretization in the spatial dimension  $x$  and the velocity dimension  $v$  holds a system of  $KJ$  ODE's in time in 1D and  $K^3J^3$  ODE's in time in 3D. By the reduction of  $v$  we arrive at  $nJ$  ODE's in time for 1D and  $nJ^3$  ODE's in time in 3D, though we discuss the 1D case only. This example should illustrate the amount of computations saved by MOR. The mapping from  $f(x, v, t)$  to  $q(x, n, t)$  is implemented through a reduction algorithm, which also performs a remapping back to  $\tilde{f}(x, v, t)$ , such that the distance  $\|f - \tilde{f}\|$  is small.

To sum up, the idea that every high dimensional dynamical-state space  $W$ , also called solution manifold, which in our case is  $f(x, v, t) \in W$ , can be mapped onto a state-space i.e.  $q_n(\mu_i)$  of lower dimension, is exploited within MOR [15]. Here  $i$  gives the number of variables that are left after the reduction and  $n$  beeing the number of so called intrinsic variables. The state space of lower dimension is called the intrinsic solution manifold  $V$  with  $q_n(\mu_i) \in V$  [5].



(a) Evolving the BGK-model in the Sod shock tube in time is generated through evaluating the FOM in space  $x$ , velocity  $v$  and time  $t$ , which yields the solution  $f(x, v, t)$ . The operator  $Q$  is here the FOM described in chapter 2.

(b) Evolving the ROM of the BGK-model in the Sod shock tube in time through evaluating the ROM at  $n$  and  $\mu_i$ , yields an approximation to the FOM solution  $\tilde{f}$ . The operator  $Q$  is here either POD or a neural network described in chapter 3.

Figure 4.1: Outline of the correlation between the FOM solution and the approximation obtained from the ROM.

MOR is partitioned into two successive phases called the *offline* - and the *online* phase. During the offline phase *snapshots* of a dynamical-system are generated through experiments or simulations of the full order model (FOM). The snapshots  $U = \{u(t_1), \dots, u(t_n)\}$  are created once, each representing one moment in time of the dynamical system. Thus in our case one needs a snapshot database of solutions  $f(x, v, t)$  of the BKG-model in the Sod shock tube. Next a mapping  $q_n$  is constructed such that  $\tilde{u} = q_n(u)$ , for which  $u(t_i) \approx \tilde{u}(t_i)$ , reducing the dimensionality of the FOM solution as outlined before. During the online phase the reduced order model is evaluated and the error is estimated by eg.  $\|u(t) - \tilde{u}(t)\|$ . Therefore the online phase may be described as a stage of independence from the full order model. Following [15] and [5], the success when building a ROM through linear reduction methods like the POD section 3.1, depends on a rapidly decaying Kolmogorov N-width. In particular advection dominated problems exhibit a slow decay of the Kolmogorov N-width. Thus yielding a need for non-linear methods like autoencoders described in section 3.2. The Kolmogorov N-width is given by

$$d_V(W) := \sup_{f \in W} \inf_{\tilde{f} \in V} \|f - \tilde{f}\| \quad (4.1)$$

and gives the worst best approximation error for elements of  $W$ . The convergence behaviour of the Kolmogorov N-width for advection dominated problems, especially when jump conditions are involved as in the Sod shock tube chapter 2, decays with

$$d_n(W) \leq \frac{1}{2} n^{-1/2}, \quad (4.2)$$

where  $n$  denotes the number of RB or intrinsic variables. Note that hereafter we will solely use the term intrinsic variables. The relevance of this contribution, using non-linear methods for MOR in rarefied gas dynamics, was already implied in the beginning and now takes shape. The BKG-model in the sod shock tube describes an advection problem with jump discontinuities chapter 2. Thus linear reduction methods should fail. We will see in the following, that this is only partly true, because the advection character as well as the formation of sharp shock fronts heavily depends on the rarefaction of the gas as depicted in chapter 2.

## 4.1 Offline Phase

### 4.1.1 Full order BGK-model

The FOM is the 1D BGK-model in the Sod shock tube for two levels of rarefaction, gratefully provided by Julian Köllemeier and the Departement of Mathematics at the RWTH Aachen. The Sod shock tube is discretized in space  $x$  with 100 nodes, in velocity  $v$  with 40 nodes for 25 time steps in time  $t$ , as presented in table 4.1 and fig. 2.2. One is a slip flow with  $Kn = 0.01$ , hereafter will be referred to as **R**. For this kind of rarefaction level the common Navier-Stokes equations are not valid anymore. The other is a continuum flow with  $Kn = 0.00001$  for which the Navier-Stokes equations can be utilized. A detailed description can be found in chapter 2. Hereafter the continuum flow will be referred to as **H**. Note that both **H** and **R** are three dimensional tensors containing  $f(x, v, t)$ .

The reduction algorithms, introduced in chapter 3, require a distinct reshaping of the input data before they can be used. The preprocessed matrix for the FCNN,  $\Pi_{\text{FCNN}}$ , is shown in eq. (4.3). Each row of  $\Pi_{\text{FCNN}}$  contains a sample shown to the FCNN. In turn we aquire  $x_i t_i = 5000$  samples.

Table 4.1: Problem setup for the BGK model in the Sod shock tube. The diaphragm is positioned at  $x_d = 0.5025$ . For  $x < x_d$  the gas is present, for  $x \geq x_d$  particles are absent.

Variable	Number of nodes $i$	Domain extension	Step size (uniform)
$x$	200	[0.0025,0.9975]	0.00499
$v$	40	[-10,10]	0.05128
$t$	25	[0,0.12]	0.005

The preprocessed matrix for the CNN,  $\Pi_{\text{CNN}}$ , is shown in eq. (4.4). We obtain  $v_i = 40$  samples, each containing a matrix  $\pi_{\text{CNN}}^{x_i \times t_i}$  holding the information about  $f(x, t)$  for a fixed point in  $v$ . POD uses the *mae* matrix or its transposition.

$$\Pi_{\text{FCNN}} = \begin{bmatrix} f(v_1, t_1, x_1) & \cdots & f(v_n, t_1, x_1) \\ f(v_1, t_1, x_2) & \cdots & f(v_n, t_1, x_2) \\ \vdots & \vdots & \vdots \\ f(v_1, t_1, x_n) & \cdots & f(v_n, t_1, x_n) \\ f(v_1, t_2, x_1) & \cdots & f(v_n, t_2, x_1) \\ \vdots & \vdots & \vdots \\ f(v_1, t_n, x_n) & \cdots & f(v_n, t_n, x_n) \end{bmatrix} \quad (4.3) \quad \Pi_{\text{CNN}} = \begin{bmatrix} n_{\text{Filters}}, & f(v_1, \mathbf{t}, \mathbf{x}) \\ n_{\text{Filters}}, & f(v_2, \mathbf{t}, \mathbf{x}) \\ \vdots & \vdots \\ n_{\text{Filters}}, & f(v_n, \mathbf{t}, \mathbf{x}) \end{bmatrix} \quad (4.4)$$

In the following a distinction between  $\Pi_{\text{CNN}}$  and  $\Pi_{\text{FCNN}}$  is omitted, when referring to the preprocessed matrices. However a distinction between the levels of rarefaction, namely  $\mathbf{H}$  and  $\mathbf{R}$ , will be utilized as  $\Pi_h$  for the former and  $\Pi_r$  for the latter.

#### 4.1.2 Constructing the mapping

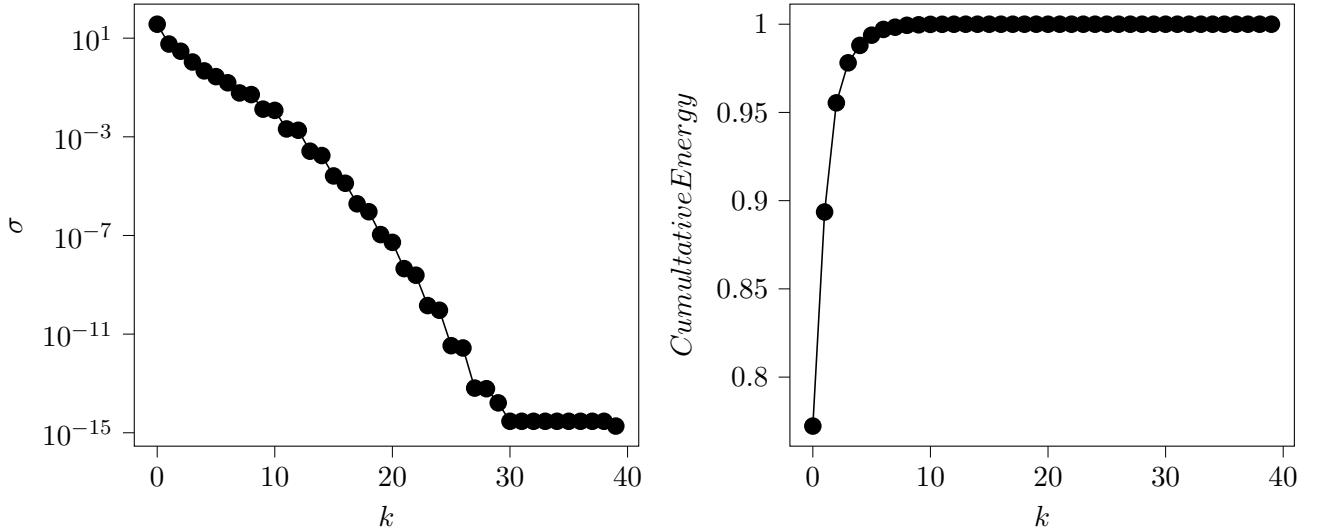


Figure 4.2: Singular values  $\sigma$  over  $k$  number of singular values left and *cumulative energy* over  $k$  right for the gas flow in the rarefied regime.

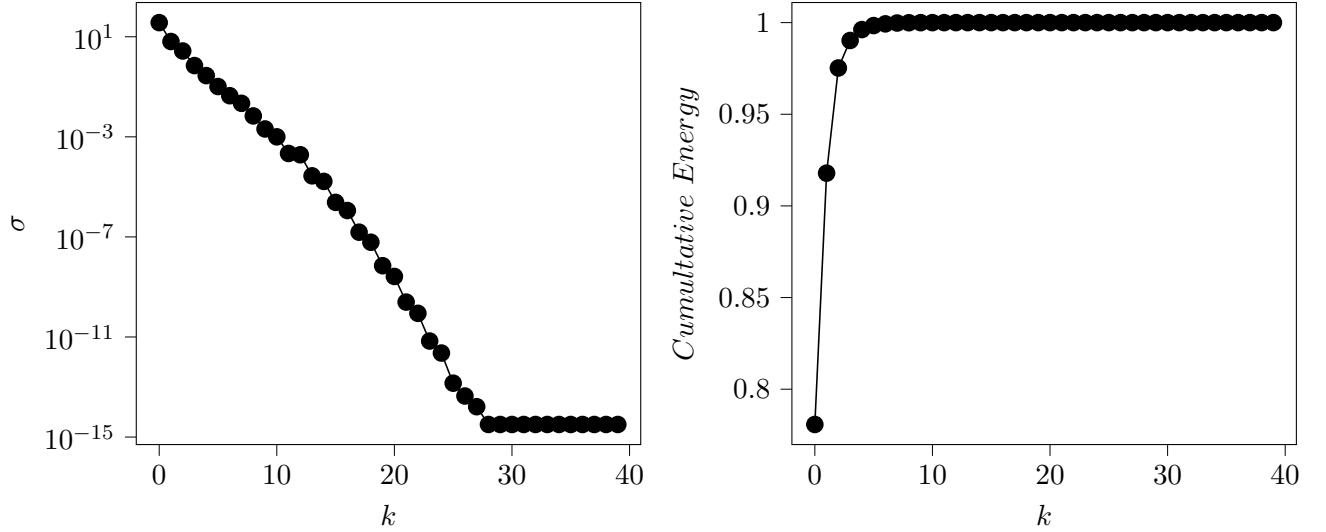


Figure 4.3: Singular values  $\sigma$  over  $k$  number of singular values left and *cumulative energy* over  $k$  right for the gas flow in the hydrodynamic regime.

Intrinsic variables	3	4	5	6	7	8	9	10
Error	0.0327	0.0153	0.0087	0.0046	0.0021	0.0014	0.0005	0.0003

Table 4.2: L2-Error for different numbers of intrinsic variables for the rarefied gas flow. Calculations with two intrinsic variables are also performed, but not shown here because two intrinsic variables is considered trivial.

Intrinsic variables	3	4	5	6	7	8	9	10
Error	0.0205	0.0081	0.0030	0.0013	0.0006	0.0002	$6.2e^{-5}$	$2.7e^{-5}$

Table 4.3: L2-Error for different numbers of intrinsic variables for the hydrodynamic gas flow. Calculations with two intrinsic variables are also performed, but not shown here because two intrinsic variables is considered trivial.

### 4.1.3 Online Phase

loremipsu

### 4.1.4 Reduced Order Model

The compression of the input data  $y_0$  yields a code  $C \in \mathbb{R}^{ix5000}$ , composed of the intrinsic variables  $c_i$ . The index  $i$  corresponds to the  $i$ -th intrinsic variable whereas their number is given by the input data. Each of them describes the transport of a discontinuity as seen in ???. Hence the expolitability of the code in terms of constructing a ROM is not provided. On that account the

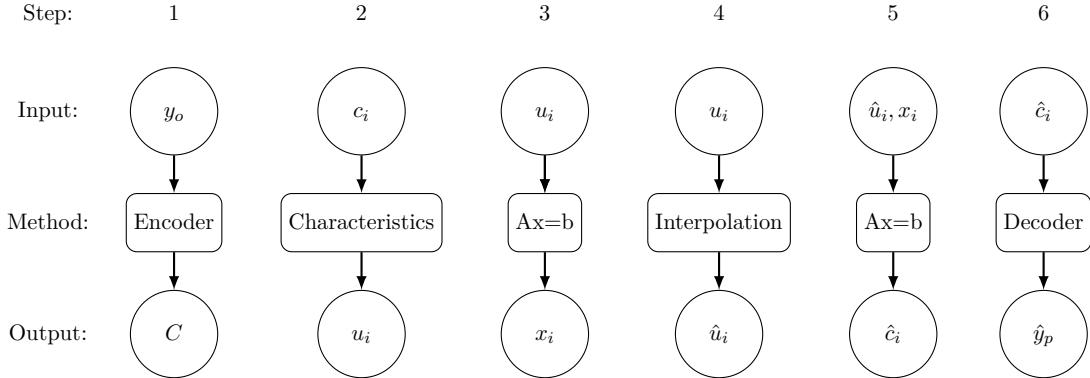


Figure 4.4: This figure shows the steps for obtaining a reduced order model (ROM). Decoder and Encoder need to be used after training. In step one  $y_o$  is the original input data,  $C$  is the Code. In step two  $c_i$  is the  $i$ -th intrinsic variable and  $u_i$  the corresponding characteristic. The eigenvalue problem in step 3 outputs  $x_i$  the eigenvector of A, a diagonal matrix composed of  $u_i$  and b is the corresponding  $i$ -th intrinsic variable  $c_i$ . In step 4  $\hat{u}_i$  is the interpolated vector to  $u_i$ . Step 5 solves the linear equation for the diagonal matrix A composed of  $\hat{u}_i$  times the eigenvector  $x_i$  of the eigenvalueproblem in step 3. The output is  $\hat{c}_i$  the  $i$ -th intrinsic variable corresponding to  $\hat{u}_i$  the  $i$ -th interpolated characteristic.

method of characteristics [16] provides a means to bypass this shortage. It is necessary for  $c_i(x, t)$  to satisfy the conservative condition eq. (4.5) and the transport equation eq. (4.6).

$$\frac{d}{dt} \int c_i dx = \frac{d}{dt} f_i = \text{const.} \quad (4.5) \qquad \frac{\partial}{\partial t} c_i + \frac{\partial}{\partial x} f_i = 0 \quad (4.6) \quad \text{The}$$

characteristics  $u_i$  describe the constant transport velocities for each variable  $c_i$  calculated using eq. (4.7). Subsequently enabling the usage of a simple polynomial interpolation of any degree. Furthermore a linear mapping  $A_i x_i = c_i$  can be applied for the reconstruction of interpolated code variables  $\hat{c}_i$ . Figure 4.4 depicts this approach in detail.

Questions concerning the capacity of this ROM, e.g. how many samples  $\hat{n}_t$  are needed to reconstruct  $n_t$  timestamps, are analysed in chapter 5.

$$u_i = \frac{f_i(c_i^-) - f_i(c_i^+)}{c_i^- - c_i^+} \quad (4.7)$$

## 5 Results

---

During the offline phase solutions obtained from the FOM, in this case  $\mathbf{H}$  and  $\mathbf{R}$ , are reduced to their intrinsic dimension, as discussed in chapter 4. From the thereby obtained intrinsic variables which define a reduced basis namely  $\mathbf{h}$  and  $\mathbf{r}$ , the solution can be reconstructed yielding a loss of information. Hence, before building a ROM, the reconstruction of  $\mathbf{H}$  and  $\mathbf{R}$  needs to be evaluated. If the reduction and subsequent reconstruction fails to sustain "important" information, then the reduction algorithm is not suited for building a ROM. What these "important" qualities for the BGK model in the test case are and how they can be measured, as well as methods to compare the FOM solution against its reconstruction is discussed in this section. For the dimensionality reduction the proper orthogonal decomposition (POD), a fully connected neural network (FCNN) and a convolutional neural network (CNN) is used.

A first measure of how much information gets lost is obtained through one of our performance metrics during neural network training: the MSE over the validation set. Yet, this metric solely applies to neural networks. To be able to compare the POD to its neural network counterparts another metric is in use: the error over the  $L_2$ -Norm here referred to as the  $L_2$ -Error, already introduced in chapter 3. In table 5.1, the  $L_2$ -Error for all three algorithms on both input data is provided. The FCNN performs best out of the three both for  $\mathbf{H}$  and  $\mathbf{R}$  and even drops about one

Table 5.1: Comparison of the  $L_2$ -Error over the reconstructions obtained by POD, FCNN and CNN. The CNN was trained using the k-fold algorithm, therefore the mean  $\mu$  over  $k = 5$  folds, marked with superscript asterix, is given. Variance  $\sigma^2$  for  $\mathbf{H}$  is  $\sigma_{\mathbf{H}}^2 = 2.25 \times 10^{-5}$  and for  $\mathbf{R}$  is  $\sigma_{\mathbf{R}}^2 = 3.61 \times 10^{-5}$ . Parenthesised values are obtained from the best fold.

Algorithm	$L_2$ -Error for $\mathbf{H}$	$L_2$ -Error for $\mathbf{R}$
POD	0.0205	0.0087
FCNN	0.0017	0.0019
CNN	0.0142* (0.0095)	0.0178* (0.0097)

decade compared to POD and CNN for  $\mathbf{H}$ . Slightly better is the result for  $\mathbf{H}$  than for  $\mathbf{R}$  using both neural network designs. Yet the performance of FCNN and CNN doesn't seem to change a lot with changing  $\mathbf{H}$  and  $\mathbf{R}$ . In contrast, the POD performs better for  $\mathbf{R}$  than for  $\mathbf{H}$ . The reasons will be discussed hereafter. Note that the CNN is trained with the k-fold algorithm due to a lack of training samples as suggested in [6]. The k-fold algorithm is provided in appendix B. The mean  $\mu$  over all folds gives an estimate of the models performance. Nonetheless the best performing fold is used in the subsequent analysis, yielding a better performance of the CNN.

Next a step is taken in attempting to estimate the intrinsic dimension  $p^*$  for  $\mathbf{R}$ , which is not set as

discussed in chapter 2. To this end the number of intrinsic variables  $p^*$  is varied for POD, FCNN and CNN over  $p^* \in \{1, 2, 4, 8, 16, 32\}$  with both  $\mathbf{H}$  and  $\mathbf{R}$ . Note that the neural networks needed to be trained again for these experiments and that by changing  $p^*$  i.e. widening the bottleneck layer, a gain or loss of capacity occurs which can be connected to stability during training, see chapter 3 and [6]. The stability of the POD, in contrast, is not altered when changing the value of  $p^*$ . Shown in fig. 5.1 is the outcome of said experiments. The design for fig. 5.1 is taken from [5].

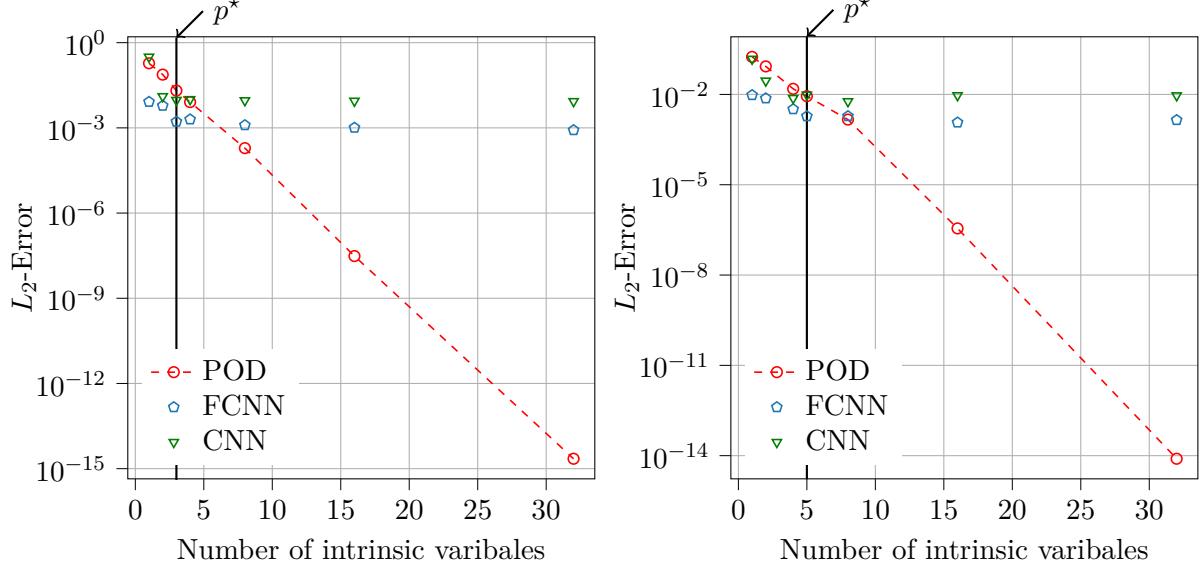


Figure 5.1: Variation of intrinsic variables over the L<sub>2</sub>-Error for POD, FCNN and CNN. Results for  $\mathbf{H}$  are displayed on the left and for  $\mathbf{R}$  on the right.

The cumulative energy of the POD modes i.e.  $p^*$  in chapter 4 fig. 4.3 and fig. 4.2 already revealed the performance of the POD with increasing  $p^*$ . The loss of information when applying the POD goes linearly to zero with increasing  $p^*$  which is not surprising when consulting the *Eckard-Young Theorem* provided in eq. (3.1) taken from [3]. The left plot of fig. 5.1 displays the results for  $\mathbf{H}$  with  $p^* = 3$  the known dimension of the intrinsic solution manifold for a simple gas which can be viewed as a continuum, emphasized with a black line. Here a sharp drop in the L<sub>2</sub>-Error can be observed for both neural networks until  $p^* = 3$ . Afterwards a stagnation is observed for CNN. FCNN continues to improve slightly. However the logically deducted and by POD proposed value of  $p^* = 3$  can be confirmed.

Moving forward to measuring the size of the intrinsic solution manifold  $p^*$  for  $\mathbf{R}$  we look at the right plot of fig. 5.1. Again, emphasized with a black line, is the from POD proposed value for  $p^*$  with  $p^* = 5$ . See again chapter 4 for details. A drop in the L<sub>2</sub>-Error can be observed for both neural networks with increasing  $p^*$  until they reach  $p^* = 5$ . Note, the CNN already arrives there with  $p^* = 4$ . The continued widening of the bottleneck layer results in an slightly increased L<sub>2</sub>-Error for both neural networks. Overfitting due to increased capacity of the model is the reason here. The discrepancy between  $p^* = 4$  and  $p^* = 5$  in the CNN and FCNN occurs since neural networks are capable of shift missing nodes i.e. weights and biases in the bottleneck layer to the weights and biases of other layers. Hence a determination of  $p^*$  by one variable is not exact with this method. As a backbone for determining  $p^*$  suits the FCNN, because of it's stable training, as seen

in appendix A, and the cumulative energy of the POD modes, as seen in chapter 4.

The "important" qualities, mentioned in the beginning, become visible when looking at the worst reconstructions. Those can be determined with the L<sub>2</sub>-Error over time for all three models as seen in fig. 5.2. For POD and the CNN the last timestep at  $t = 0.12s$  in both cases **H** and **R** is the most

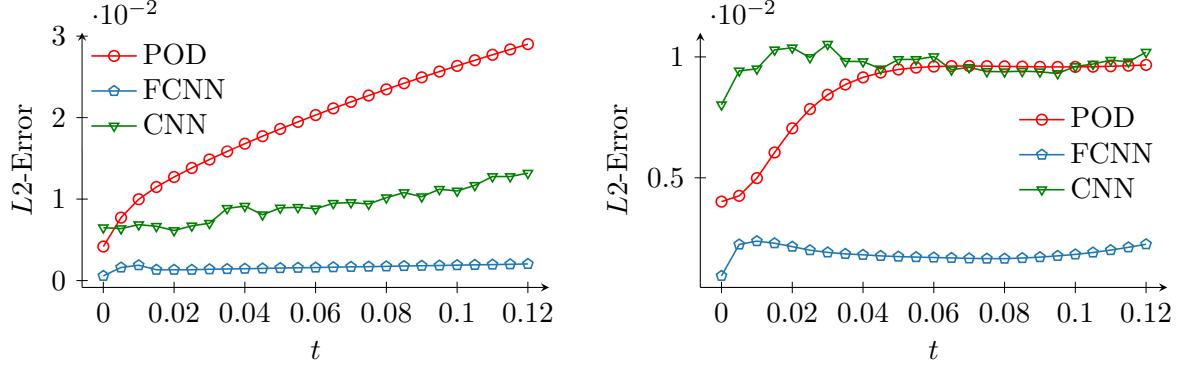


Figure 5.2: Relative Error over time for POD, FCNN and CNN. Results for **H** are displayed on the left, the results for **R** are displayed on the right.

rich in the L<sub>2</sub>-Error. While performing best, the FCNN in contrast has troubles in the beginning around  $t = 0.01s$  for **H** and **R**. A detailed comparison of the models, reconstructions of  $f(x, v)$  at  $t = 0.12s$  and  $x \in [75cm, 150cm]$  are given in fig. 5.3.

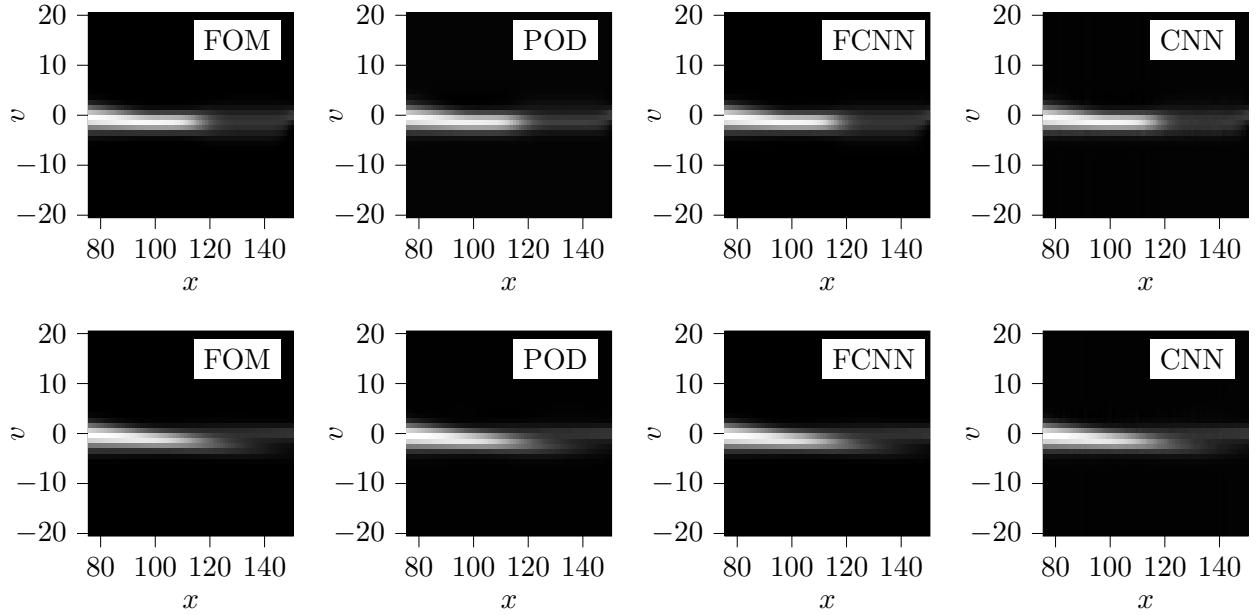


Figure 5.3: Comparison of FOM with three reconstructions obtained from POD, FCNN and CNN. The BGK-model in Sod's shock tube at time  $t = 0.12s$  and  $x \in [75cm, 150cm]$  of the tube, is most difficult to reconstruct by the aforementioned algorithms. Case **H** is displayed in the top row, **R** in the bottom row.

A presentation of the FOM in  $f(x, v)$  was already introduced in chapter 2. For clarity,  $f(x_0, v)$  is a probability distribution for a gas having a velocity  $v$  at point  $x_0$  in space at one moment  $t_i$  in time. Top row in fig. 5.3 show reconstructions for **H** and the bottom row for **R** with FOM solution as paradigm. Starting with **H** one observes, that a restored  $f(x, v)$  from  $x = 120$ , the point around which dilution initiates, gets defective for POD and the CNN. Here the probability distribution is thinner as the original with POD and smeared around the borders with the CNN. This in turn leads to errors in velocities gas particles can have once passing  $x = 120$ . In contrast the FCNN reproduces the FOM solution almost exactly.

Continuing with a row further down of fig. 5.3 and therefore **R**, in place of a pronounced separation in a dense and rare region, stands bifurcation of  $f(x, v)$  into two probability distribution functions as outlined in chapter 2. POD and the FCNN reproduce the FOM solution without any visible drawback. On the other hand the CNN reproduces smeared corners as in **H**.

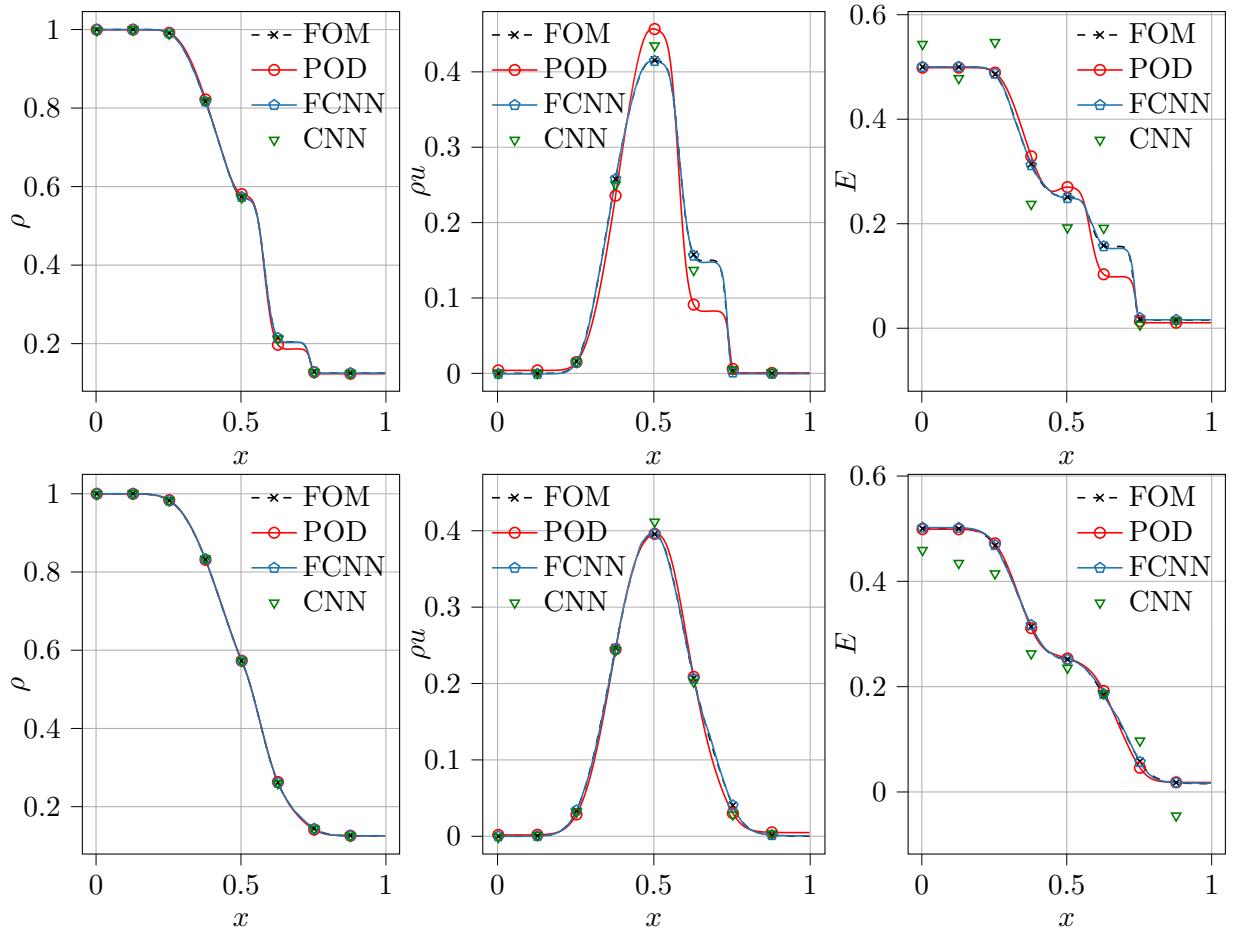


Figure 5.4: Matching of macroscopic quantities  $\rho$ ,  $\rho u$  and  $E$  reproduced by POD, FCNN, and CNN with FOM macroscopic quantities. Top row shows results for **H**, bottom row for **R**. CNN is displayed with marks only because of trembles in the signal.

Loss of information described above can unfold in severe mistakes in  $\rho$ ,  $\rho u$  and  $E$ , the macroscopic quantities, as displayed in fig. 5.4. In the following, features of the macroscopic quantities are

expressed in terms of rarefaction wave, contact discontinuity and height as well as position of the shockfront. For a detailed elaboration see chapter 2. Following the structure in the preceding figures, macroscopic quantities of  $\mathbf{H}$  are displayed in the top row and for  $\mathbf{R}$  in the bottom row of fig. 5.4. First the reproduction of the macroscopic quantities  $\rho$ ,  $\rho u$  and  $E$ , obtained by the FCNN is exact for both cases  $\mathbf{H}$  and  $\mathbf{R}$ . In particular the density  $\rho$  matches with results from the FOM exactly for the neural networks in both cases  $\mathbf{H}$  and  $\mathbf{R}$ . Second the CNN produces trembles in  $\rho u$  and especially in  $E$  which is why it's shown with marks only. Results from the CNN are similar for  $\mathbf{H}$  and  $\mathbf{R}$ . The momentum  $\rho u$  holds errors for the tail of the rarefaction wave as well as the contact discontinuity and the position and height of the shockwave. Third POD performs better on  $\mathbf{R}$ , holding only small deviations in the contact discontinuity and position and height of the shockwave for the momentum  $\rho u$  and the total energy  $E$ . The density  $\rho$  matches with the FOM solution exact. Distinct deviations from the FOM solution occur using POD on  $\mathbf{H}$ . The density  $\rho$  holds errors in the height of the shockwave. The momentum  $\rho u$  holds errors in the tail of the rarefaction wave, the contact discontinuity and the height of the shockwave. Hence in the total energy  $E$  errors occur in the rarefaction wave, especially the tail, the contact discontinuity and the height of the shockwave.

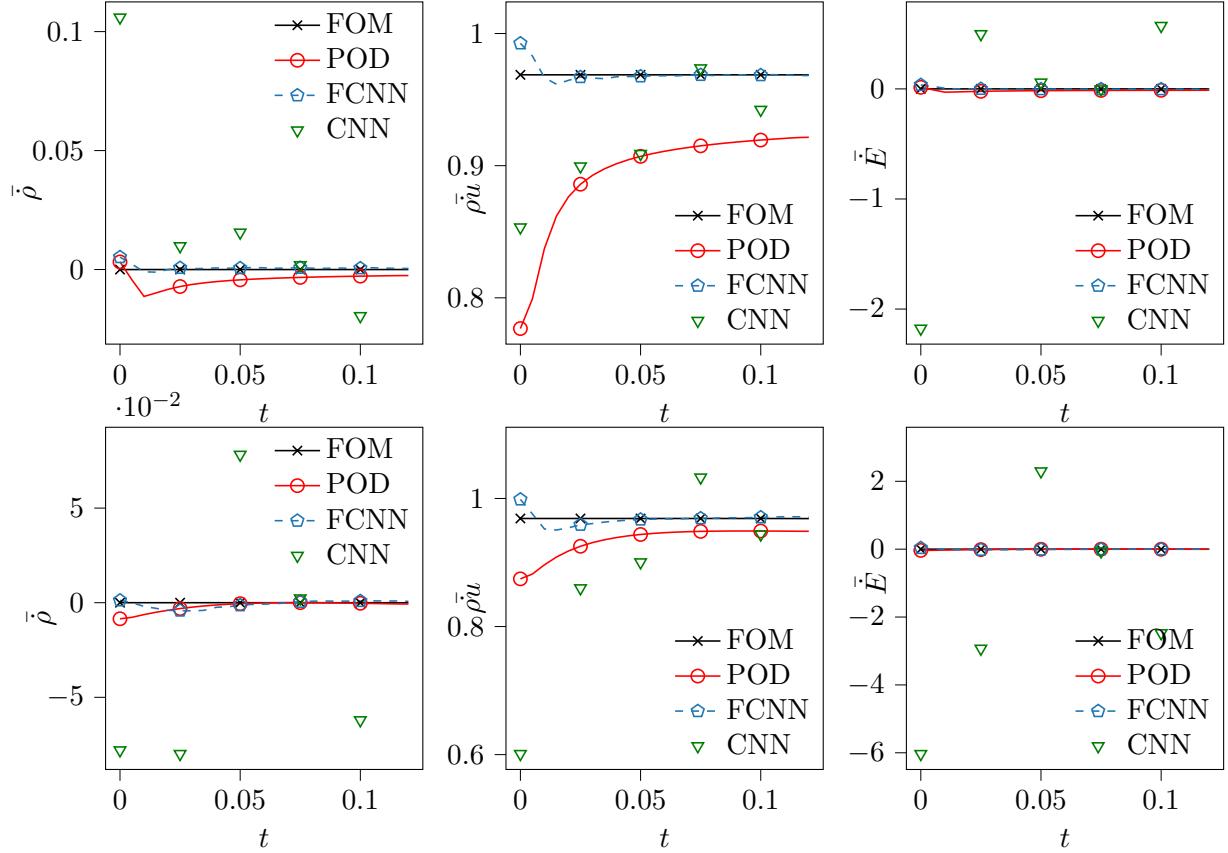


Figure 5.5: Comparison of the conservative properties of reconstructions obtained from POD, the FCNN and the CNN against the conservative properties of the FOM solution using the temporal mean.

Conservative properties of the FOM are discussed in chapter 2. Now we want to estimate if the conservation of mass, momentum and total energy could be sustained using POD, the FCNN, and the CNN. To do so, the temporal mean over the time derivative of the macroscopic quantities is employed. Figure 5.5 shows the conservation of mass, momentum and total energy over time for **H** in the top row and for **R** in the bottom row.

Conservation of mass is met using the FCNN, except for small deviations at the outset for both cases **H** and **R**. Similarly, does POD meet conservation of mass for **R**. The erroneous **H** case shows deviations from conservation with POD. Conservation of momentum meets the FOM solution after  $t = 0.03s$  using the FCNN for both cases **H** and **R**. POD conserves momentum close to the FOM solution, but deviations are similarly present for **H** and **R**. Next conservation of total energy is met for **H** and **R** using POD and the FCNN. Finally the reconstructions of the CNN do not conserve mass, momentum nor total energy. All conservative properties behave comparable to a sawtooth wave. A gain and loss of either of the quantities can be observed.

In conclusion the error over time for the CNN is disordered, showing gain and subsequent loss of information from one timestep to another. The CNN performs slightly better with **H** than with **R**. What is hidden when looking at reconstructions becomes visible when verifying over the macroscopic quantities. Reconstructions obtained from the CNN show oscillations in the momentum  $\rho u$  and the total energy  $E$ . On top of that the CNN does not meet conservation in any of the conservative properties. All of this together makes the CNN with this setup, especially the access to only 40 samples, unsuited for building a ROM. Next POD shows a noticeable increase in loss of information over time for both cases **H** and **R**. Reconstructions of the last timestep as well as the macroscopic quantities at that time reveal that the POD is unsuited for building a ROM with the **H** case. However, with **R** POD shows only slight deviations from the FOM solution. Taking conservative properties of the reconstructions obtained from POD into consideration only underlines aforementioned findings. Ultimately POD could be taken for building a ROM with **R**. Finally the the FCNN is the best performing model out of the three for both cases **H** and **R**, while the performance for **H** is slightly better than that for **R**. The error over time reveals a constant low loss. Only at the first time steps a noticeable loss of information is observed. Reconstructions of the last time step and the macroscopic quantities at that time are close to exact to the FOM solution for both cases **H** and **R**. The conservation of the macroscopic quantites only emphasize the proximity to the FOM solution. In total the FCNN is suited for building a ROM with both cases **H** and **R** and will be taken further into the online phase.

We now reached the online phase where we want to be independent from the FOM solution. A first ROM relying on pure interpolation in the intrinsic variables is performed for **H**. The intrinsic variables of the FCNN for **H** are shown in ... .

### 5.0.1 Discussion and Outlook

One reason of the lacking ability of the CNN is the small number of samples, as described in appendix B and chapter 4. The resulting trembles of the signal when calculating the macroscopic quantities is due the kernel approach of the CNN. During reconstruction the resolution of the output image is bounded by the size of the kernel, which leads to pixelation.

What we see here is the known drawback of POD. Sharp fronts and especially advection dominated problems lead to a fast decaying kogolomorov n-width. These problems need a nonlienar ansatz, as described in chapter 4.

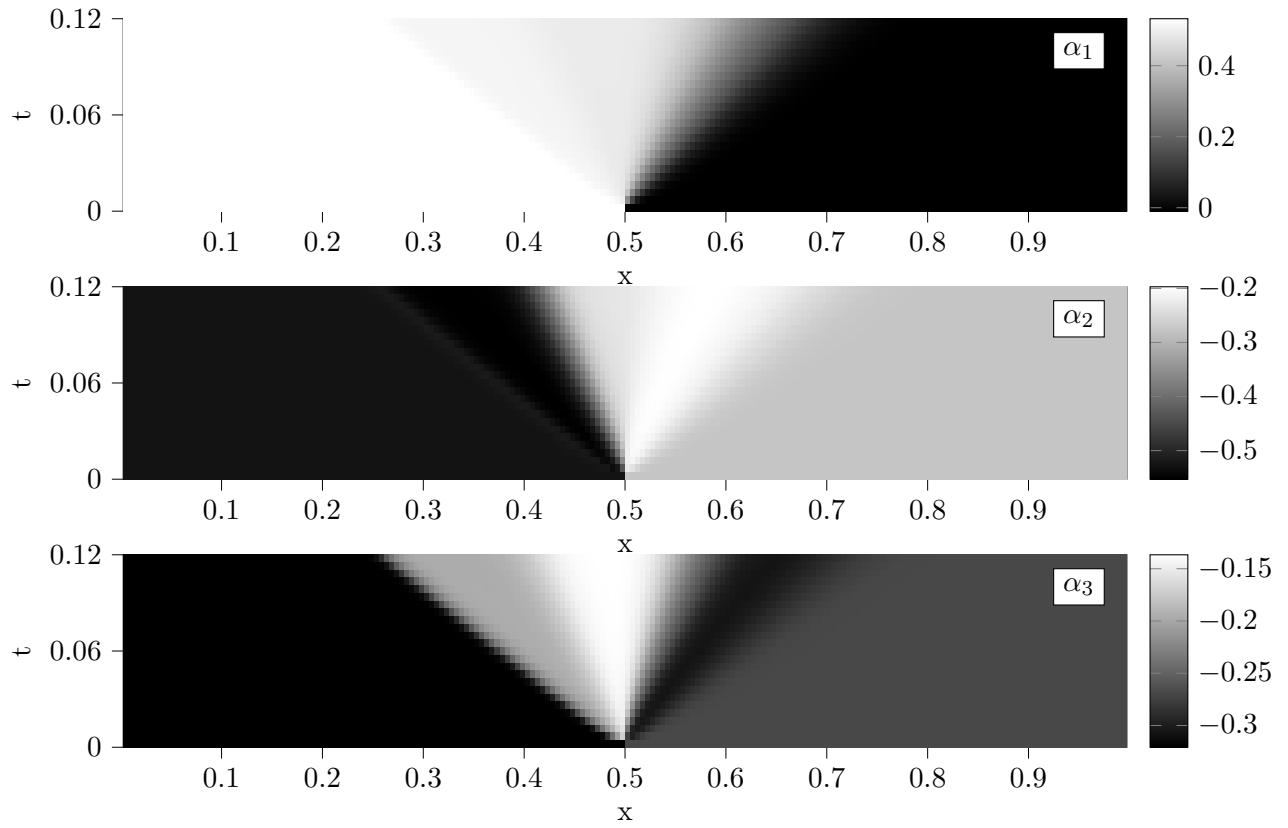


Figure 5.6:  $\alpha_1$ ,  $\alpha_2$  and  $\alpha_3$ , the reduced basis  $\mathbf{h}$  obtained from the FCNN.

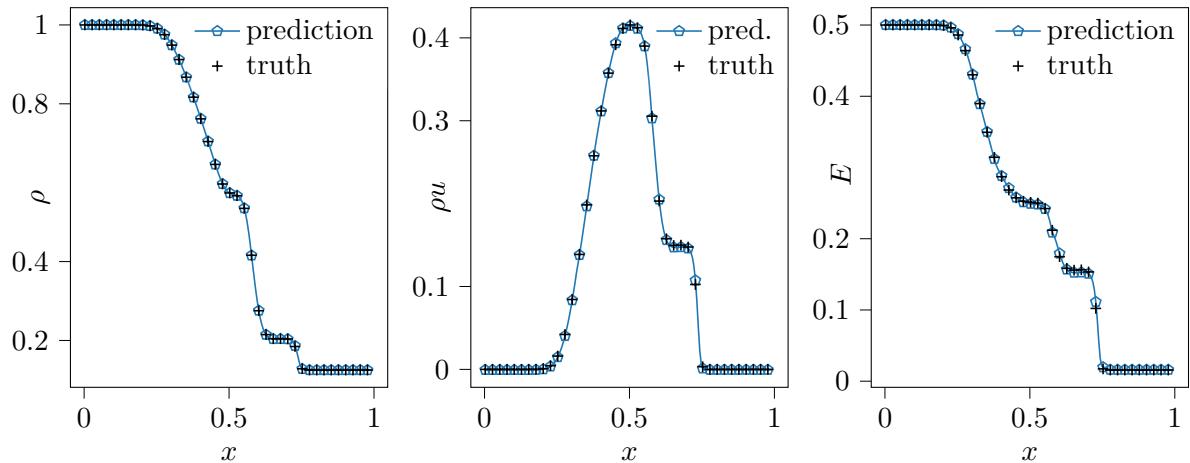


Figure 5.7: Interpolation in time for  $\mathbf{H}$  from 25 snapshots to 241 snapshots with cubic splines using the FCNN.

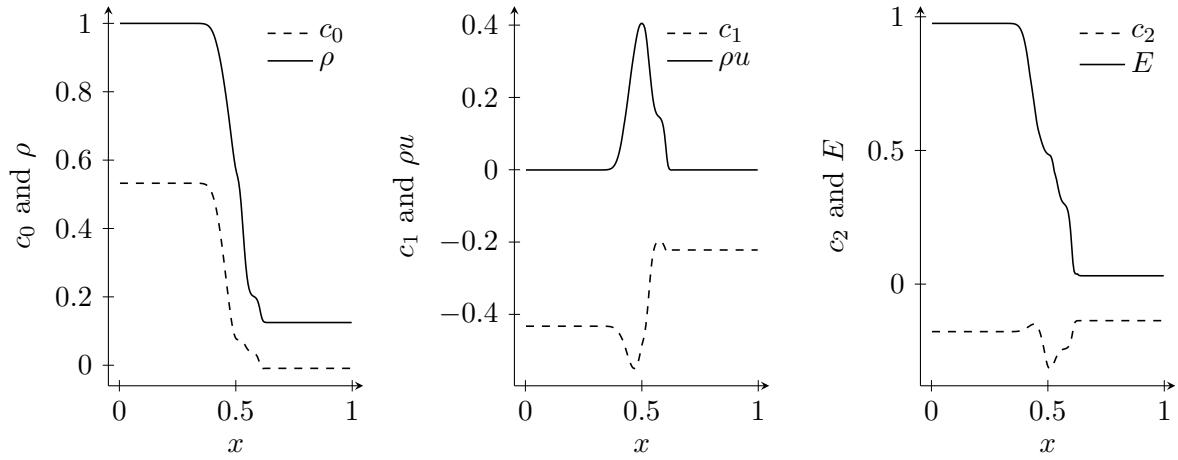


Figure 5.8: Code variables  $c_1$ ,  $c_2$  and  $c_3$  (dashed lines - -) and macroscopic quantities  $\rho$ ,  $E$ ,  $\rho u$  (full lines -) for  $t = 0.05s$ .

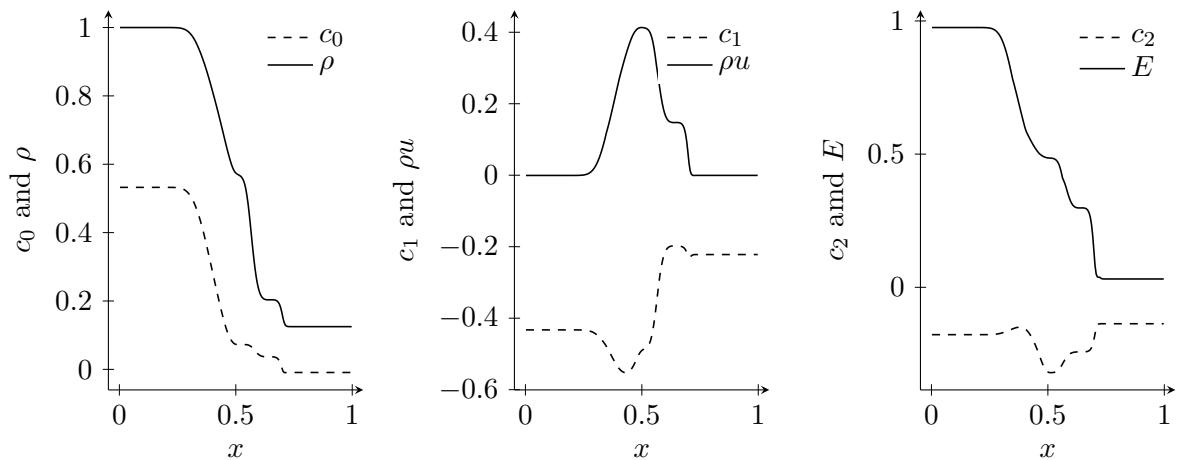


Figure 5.9: Code variables  $c_1$ ,  $c_2$  and  $c_3$  (dashed lines - -) and macroscopic quantities  $\rho$ ,  $E$ ,  $\rho u$  (full lines -) for  $t = 0.099s$ .

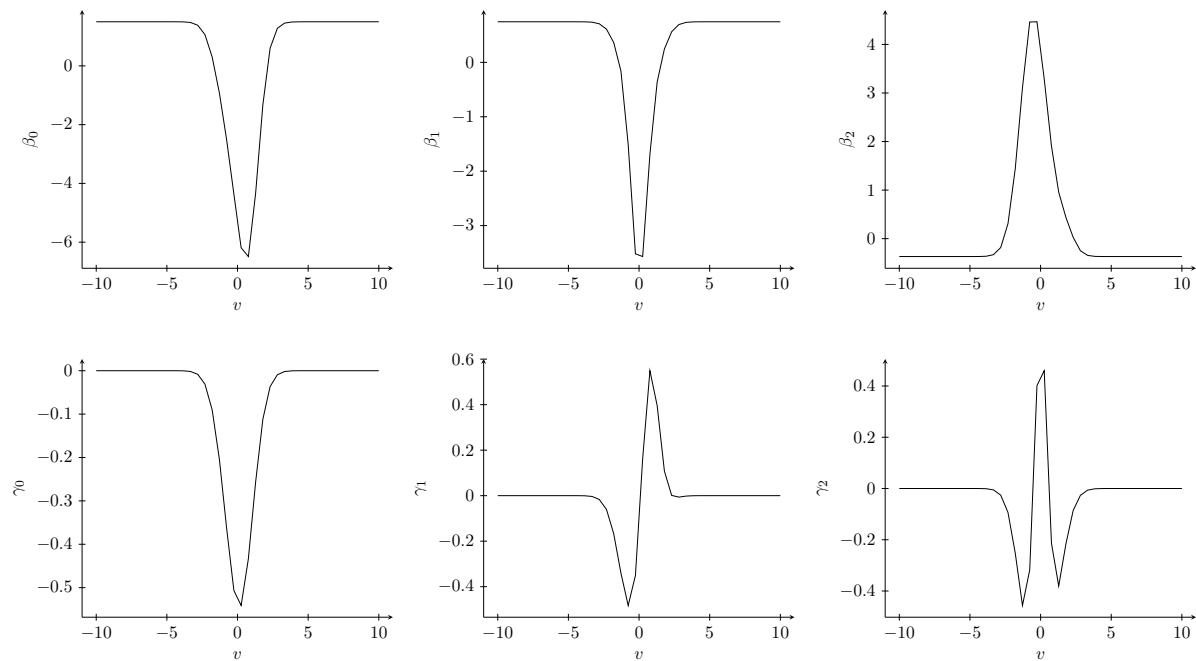


Figure 5.10: Comparison of the intrinsic variables generated by POD  $\gamma_1$ ,  $\gamma_2$  and  $\gamma_3$  with the intrinsic variables of the convolutional autoencoder  $\beta_1$ ,  $\beta_2$  and  $\beta_3$ .

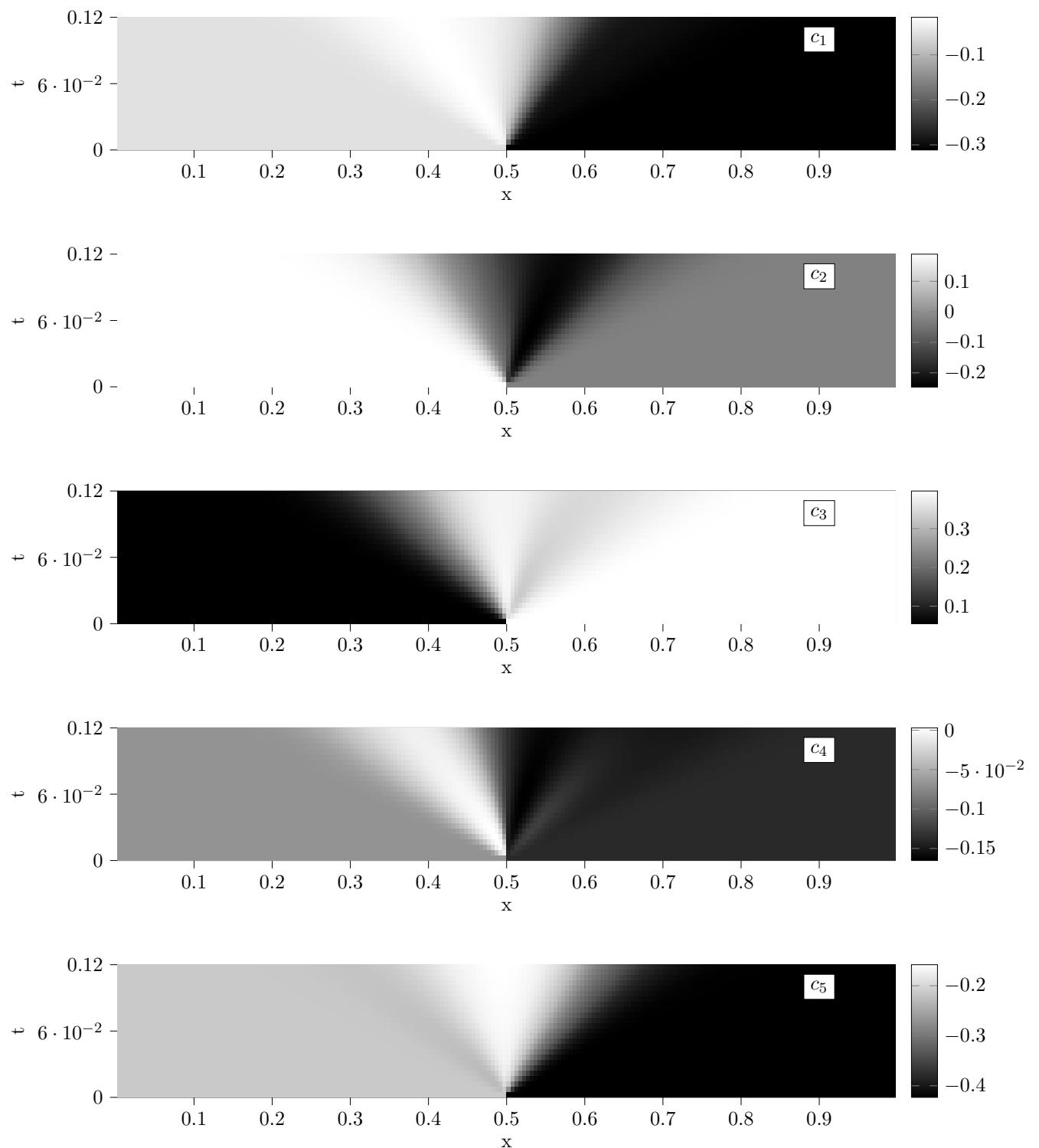


Figure 5.11: Intrinsic Variables of  $\mathbf{R}$

## **Appendix**

## **A Hyperparameters for the Fully Connected Autoencoder**

To start with a working model a guess is needed to be made about the initial design of the architecture. Here the hyperparameters already found are used as a starting point and shown in Table A.1.

To start with the number of layers or depth as described in ?? is determined, as they set a

Mini batch size	Intrinsic dimensions	Epochs	Learing rate	activation code/rest
16	3	2000	1e-4	Tanh/LeakyReLU

Table A.1: Initial hyperparameter selection

consequential part of the representational capacity of the model and therefore can initiate over- and underfitting at an early stage of the parameter search. ?? and ?? in ?? show the training and validation error for five designs shown in table A.2.

For the hydrodynamic regime a number of layers greater than four results in a slight overfitting

Number of layers	Reduction per layer
10	40 → 40 → 20 → 10 → 5 → 3
8	40 → 40 → 20 → 10 → 3
6	40 → 40 → 20 → 3
4	40 → 40 → 3
2	40 → 3

Table A.2: Initial hyperparameter selection

at an early stage of training (at around 100 epochs). Below four layers an underfitting can be observed. Hence yielding the conclusion, that four layers result in the best performing net at this early stage. Overfitting occurs only after the 1000th epoch and is less than with the other three nets that show overfitting. Note, that contrary to expected results, the train error is lower than the test error. The random shuffling during preprocessing might be taken into account here. Nonetheless solely overfitting is defined by differences between train and test loss. In Addition four layers and lower show a stable training in relation to the rest.

The analysis of the number of layers for the rarefied regime is not showing overfitting as obvious as before. The network with two layers underfits in contrast to the other nets whereas nets with more than two layers reach a similar train - and test loss. Networks with more than 4 layers, again show an unstable training compared to the net with four layers. Train and test loss show a diverging behaviour after around the 100th epoch.

In conclusion the net with four layers performs best for both the hydrodynamic and the rarefied regime.

Training duration is raised from 2000 epochs to 5000, as training did now converge completely as seen in ?? and fig. A.2.

In the following the width of the hidden layer will be analysed. There are two available hidden layers for the autoencoder. But as the architecture of the decoder mirrors that of the encoder only one parameter needs to be varied. For both the hydrodynamic and the rarefied regime five experiments are conducted, varying the hidden units from ten to fifty. Results for  $\Pi_h$  and  $\Pi_r$  are shown in table A.3. For **H** and **R** forty hidden units performs best with an error around  $1e^{-3}$  and

Hidden units	10	20	30	40	50
Error for $\Pi_h$	0.0054	0.0027	0.0036	0.0017	0.0032
Shrinkage factor for <b>H</b>	0.3	0.1	0.015	0.075	0.06
Error for $\Pi_r$	0.0078	0.0027	0.0022	0.0015	0.0025
Shrinkage factor for <b>R</b>	0.5	0.25	0.016	0.0125	0.01

Table A.3: L2-Error for different number of hidden units for  $\Pi_h$  and  $\Pi_r$ .

a shrinkage factor of 0.075 for **H** and 0.0125 for **R**. The combination of representational capacity set by the number of hidden units and shrinkage factor has an optimum here. When decreasing the number of hidden units, the representational capacity of the model also decreases. At the same time the shrinkage factor increases, therefore less information needs to be compressed in one step. The error increases with decreasing the number of hidden units from forty. This can be explained by the lower representational capacity. When increasing the number of hidden units to fifty the error also increases though, the representational capacity grows. Here the decreased shrinkage factor could be responsible for that.

Next the mini-batch size is analysed. Results are displayed in table A.4 for **H** and in table A.5 for **R**. Additionally appendix B.0.1 shows the training for both input data **H** and **R**. For both input data experiments are first conducted with mini-batch sizes in the range of : [2, 4, 8, 16, 32].

The results of the L2-Error show best performance between a mini-batch size of 8 and 16 for **H**. Likewise the results of the L2-Error for **R** show best performance between mini-batch sizes of 4 and 16. Therefore additional experiments are conducted. Three additional experiments for **H** with mini-batch sizes of: [10, 12, 24]. Two additional experiments for **R** with mini-batch sizes of: [6, 10]. The training reveals that smaller batch sizes lead to oscillating training and test errors. This can be overcome by a smaller learning rate, but also leads to increased computational time as more epochs are needed to achieve a comparable training and test loss at the last epoch. A mini-batch size of 8 is bordering the zone where the oscillations of training and test loss subside. At the same time a mini-batch size of 8 to 10 indicate also a growth of the L2-Error. In conclusion a mini-batch size of 8 is chosen to continue to work with for **H** and **R**. The oscillations which make the training instable can be battled with a lower learning rate as soon as training starts to tremble.

Together with the mini-batch sizes, activations are examined. Hence experiments with different

Batch Size	2	4	8	10	12	14	16	32
Error for $\Pi_h$	0.0011	0.0011	0.0011	0.0029	0.0018	0.0018	0.0013	0.0030

Table A.4: L2-Error for different mini-batch sizes for  $\Pi_h$ .

activation functions are performed with the initially used mini-batch size of 16. Besides the mini-

Batch Size	2	4	6	8	10	16	32
Error for $\Pi_r$	0.0014	0.0010	0.0012	0.0012	0.0014	0.0017	0.0017

Table A.5: L2-Error for different mini-batch sizes for  $\Pi_r$ .

batch size, epochs are as well taken from the initial selection of hyperparameters as 2000 epochs. Unlike the previous two examinations the selection of activation functions shows early in training if an acceptable level of performance can be achieved. This is a personal observation made throughout working on this thesis. First five activations are applied to all the four layers. Second a combination of activation functions is studied, where the intrinsic variables is activated with a different function than the remaining layers. Results can be observed in table A.6 and table A.7. Both input data

Activation function	ReLU	ELU	Tanh	SiLU	LeakyReLU
Error	0.0028	0.0019	0.0036	0.002	0.0039
Activation function	ELU/Tanh	LeakyReLU/Tanh	ELU/SiLU		
Error	0.0019	0.0017	0.0019		

Table A.6: L2-Error different activation functions and combinations for  $\mathbf{H}$ .

Activation function	ReLU	ELU	Tanh	SiLU	LeakyReLU
Error	0.0048	0.0029	0.0037	0.0134	0.0034
Activation function	ELU/Tanh	LeakyReLU/Tanh	ELU/SiLU		
Error	0.0032	0.0030	0.0036		

Table A.7: L2-Error different activation functions and combinations for  $\mathbf{R}$ .

lead to different results this time. For that reason  $\mathbf{H}$  and  $\mathbf{R}$  are studied separately. Starting with  $\mathbf{H}$  and the values of the L2-Error, it can be observed that, while ELU and SiLU stand out with the lowest value of the L2-Error, when applying the same activation function to all layers, all three coupled activations convince in the combination case. Applying different activations to the layers increases the representational capacity of the model since the model can feed on a greater variety of functions as explained in the previous subsection. The need for a great representational capacity reduces for  $\mathbf{H}$ . The reasons are the linearity of that case as described in [1]. Hence two activations (ELU and SiLU) from a similar family of functions perform well on  $\mathbf{H}$  without the need of a second activation. Still adding another activation render as good results concerning the L2-Error. When observing the train -and test loss in fig. A.5 one can notice that overfitting clusters the activations into two groups. While they all achieve a similar MSE-Loss for train -and test data only ELU and Tanh are not showing overfitting. The combination of two activation functions yields an abundance of overfitting only for the combination of LeakyReLU with Tanh. In addition LeakyReLU with Tanh reaches together with ELU in combination with SiLU the lowest train -and test loss. The slight overfitting observed with the combination ELU and SiLU give rise to choose for the final model LeakyReLU with Tanh.

To continue with  $\mathbf{R}$  the L2-Error is again observed in table A.7, ELU stands out for the single activation. ELU with Tanh and LeakyReLU with Tanh meet the lowest L2-error for the combination of activations. Next observing the training for the three activations in fig. A.6 one finds that all

three show little to no overfitting, while the combination of LeakyReLU and Tanh deviates from that slightly. From this point no clean decision can be made which activation or combination to take. Therefore another run is conducted for the three. Results are shown in table A.8. Only LeakyReLU with Tanh and ELU alone convince with a lower L2-loss. The train -and test loss does not give any hint which of the two remaining activations/ combination of activations to take concerning overfitting or a profound gradient for the train -and test loss. Hence a third run is performed taking the parameters already produced for the two models and training them for another 2000 epochs. This warm-starting of the models finally gives an answer on which one to take. The results in table A.8 show that LeakyReLU with Tanh, just as for **H**, perform well on **R** with the given model. In fig. A.7 the train and test loss during training are provided for the second and third run.

Activation function	ELU	ELU/Tanh	LeakyReLU/TanH
Error 2nd run	0.0034	0.0029	0.0025
Error 3rd run		0.0024	0.0019

Table A.8: L2-Error different activation functions and combinations for **H**.

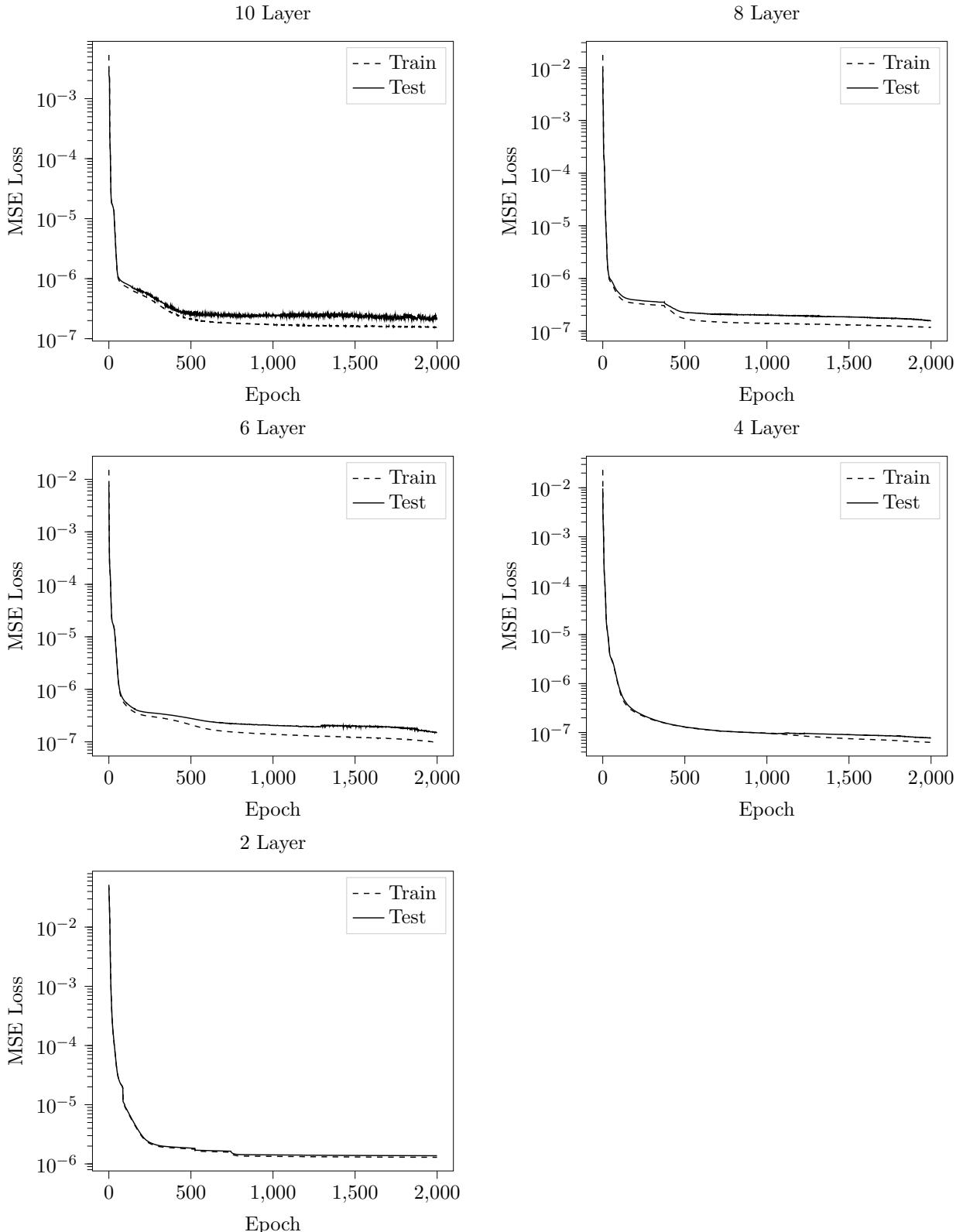


Figure A.1: Experiments with five different depth on  $\mathbf{H}$ . Training and test loss are shown over 2000 epochs.

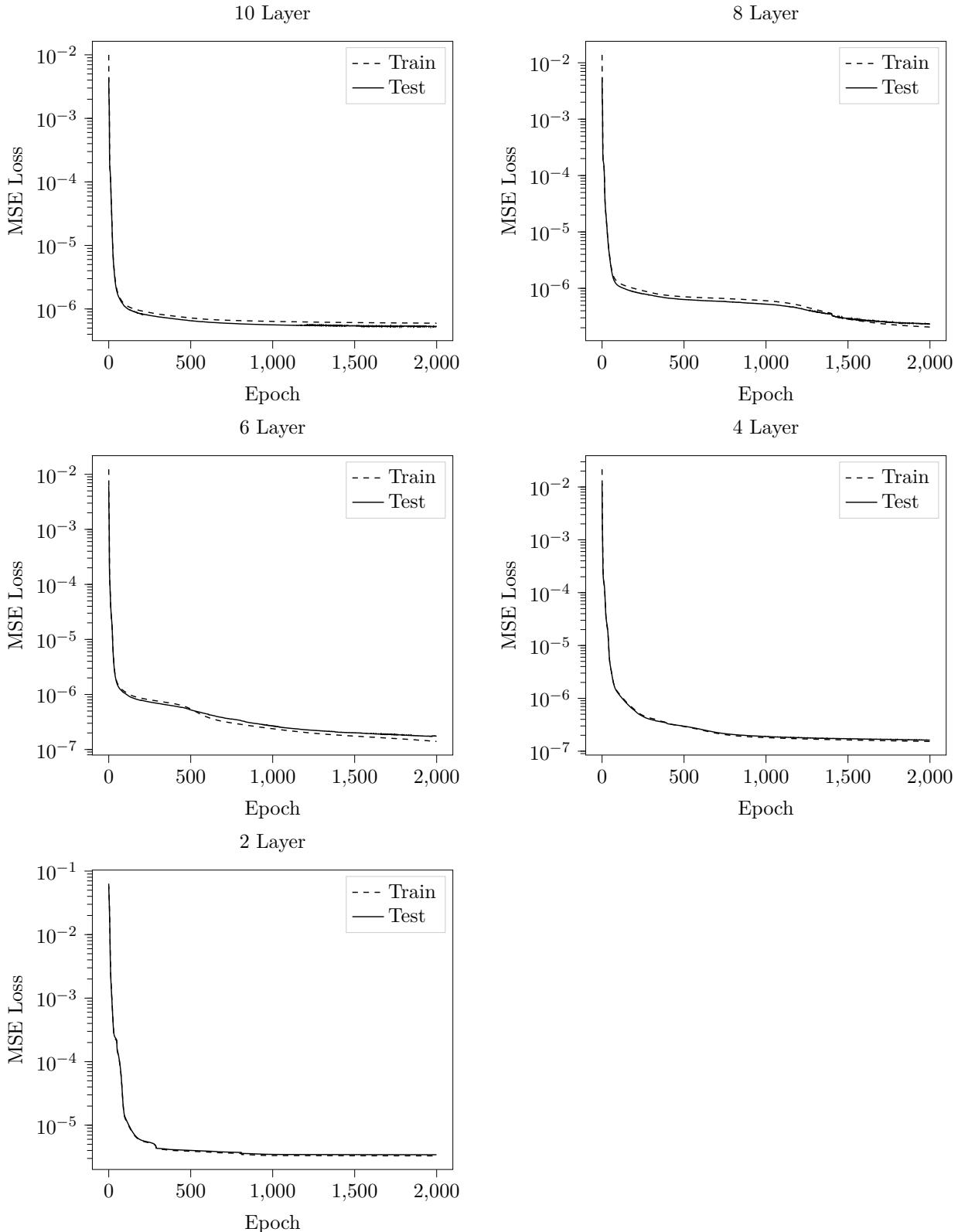


Figure A.2: Experiments with five different depth on  $\mathbf{R}$ . Training and test loss are shown over 2000 epochs.

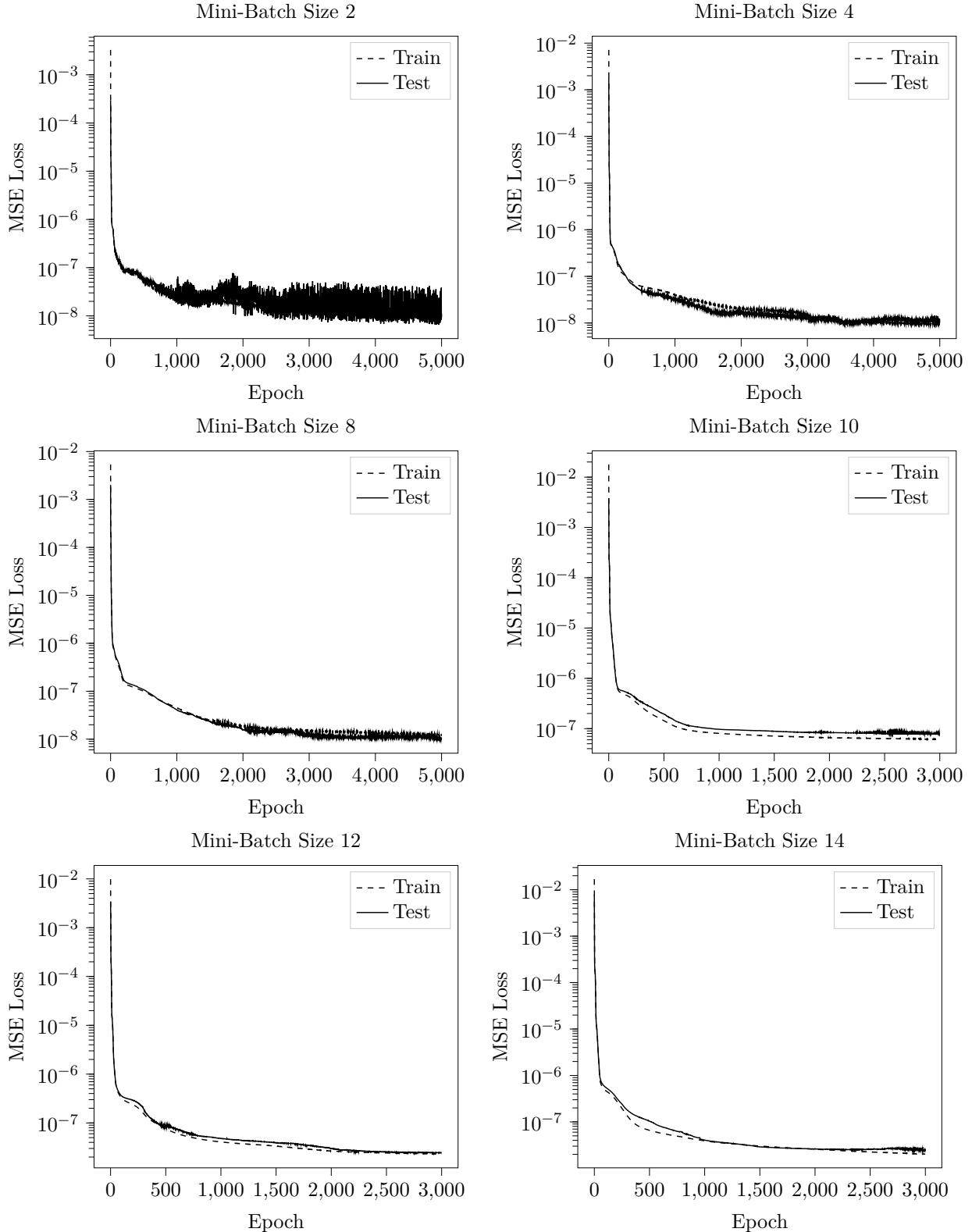


Figure A.3: Training with different mini-batch sizes for  $\mathbf{H}$ . Train -and test loss is shown over 5000 epochs.

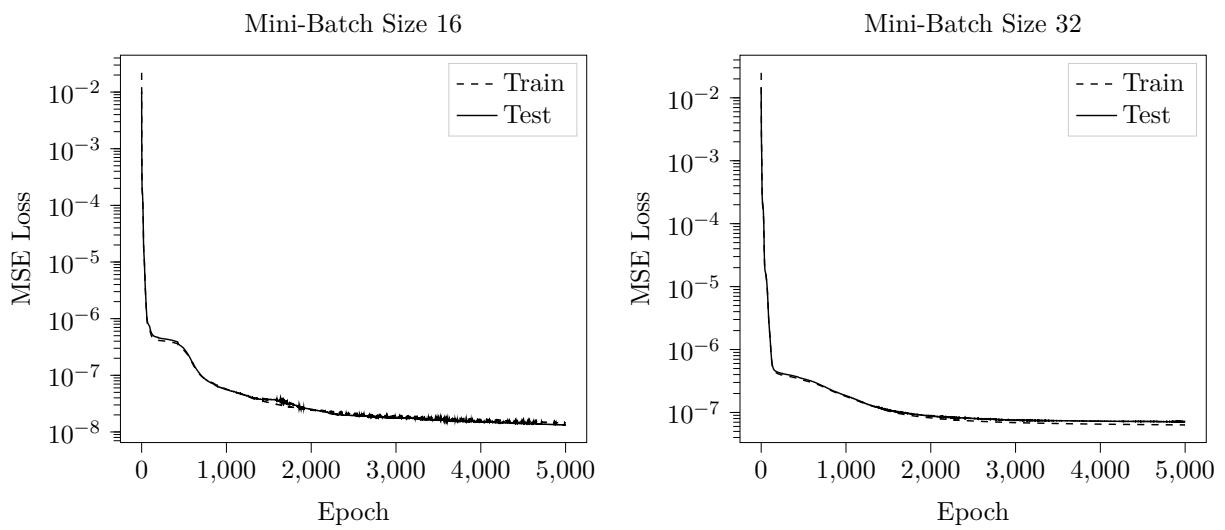


Figure A.3: Training with different mini-batch sizes for  $\mathbf{H}$ . Train -and test loss is shown over 5000 epochs.

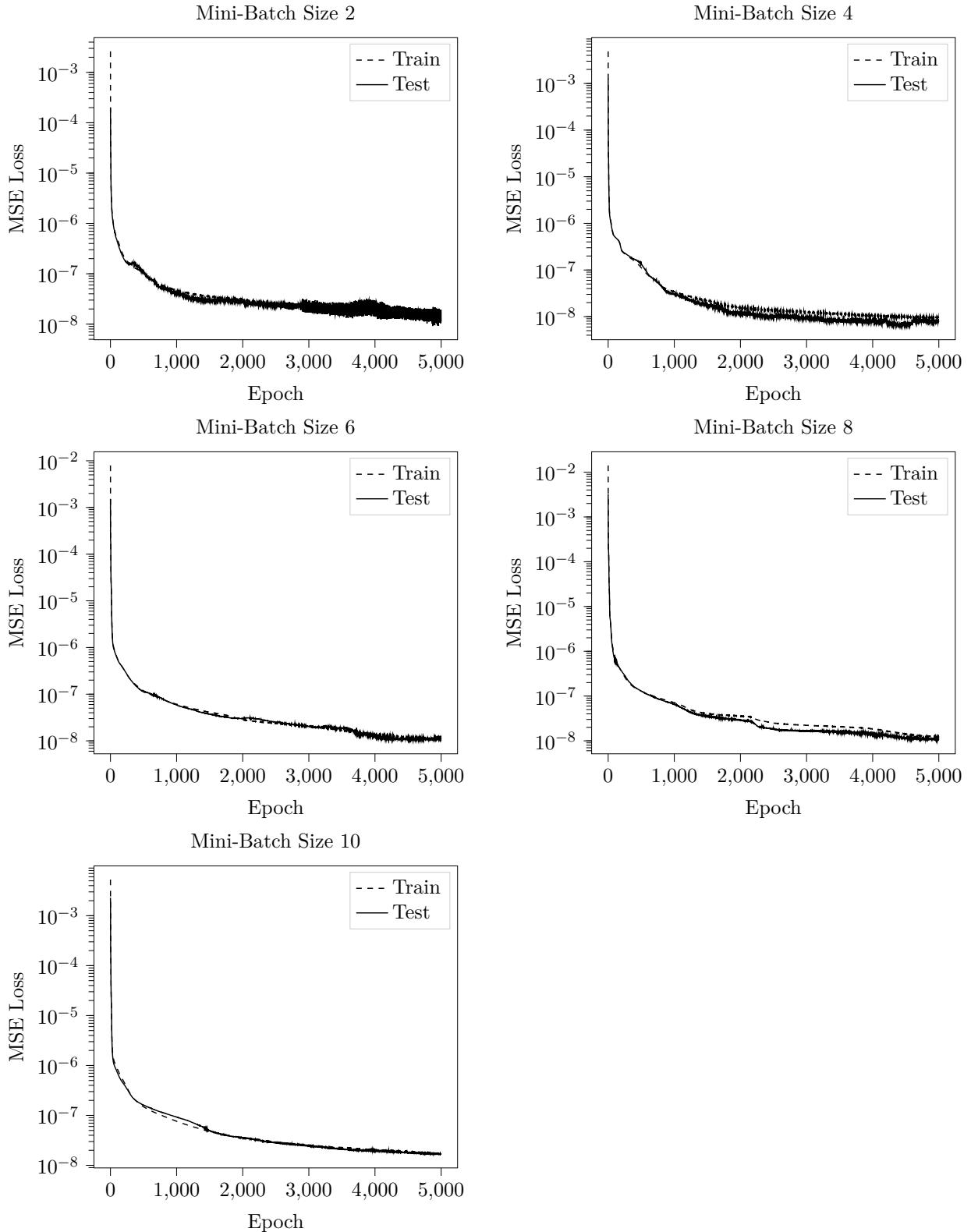


Figure A.4: Training with different mini-batch sizes for  $\mathbf{R}$ . Train -and test loss is shown over 5000 epochs.

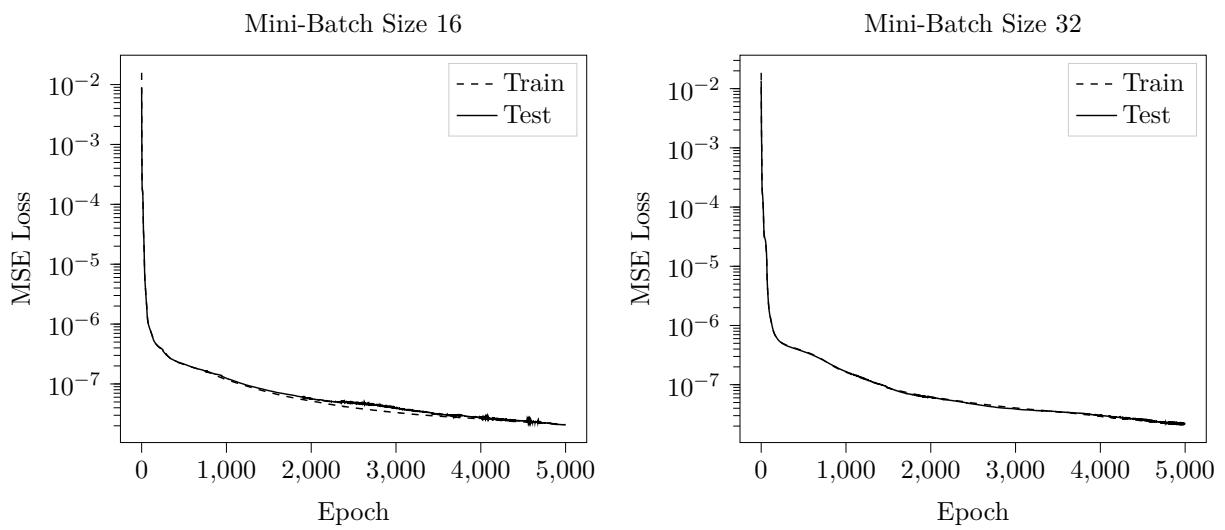


Figure A.4: Training with different mini-batch sizes for **R**. Train -and test loss is shown over 5000 epochs.

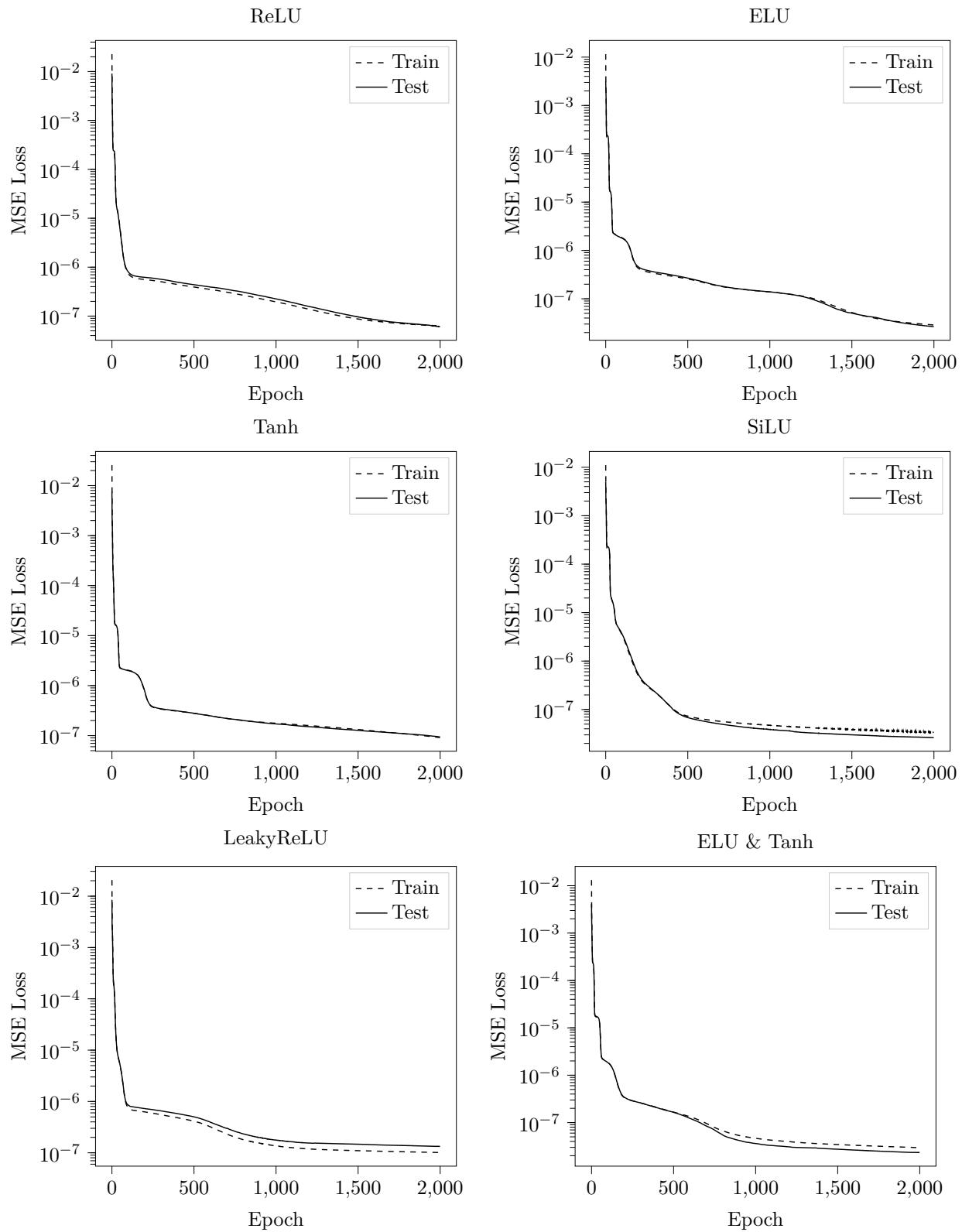


Figure A.5: balblabla

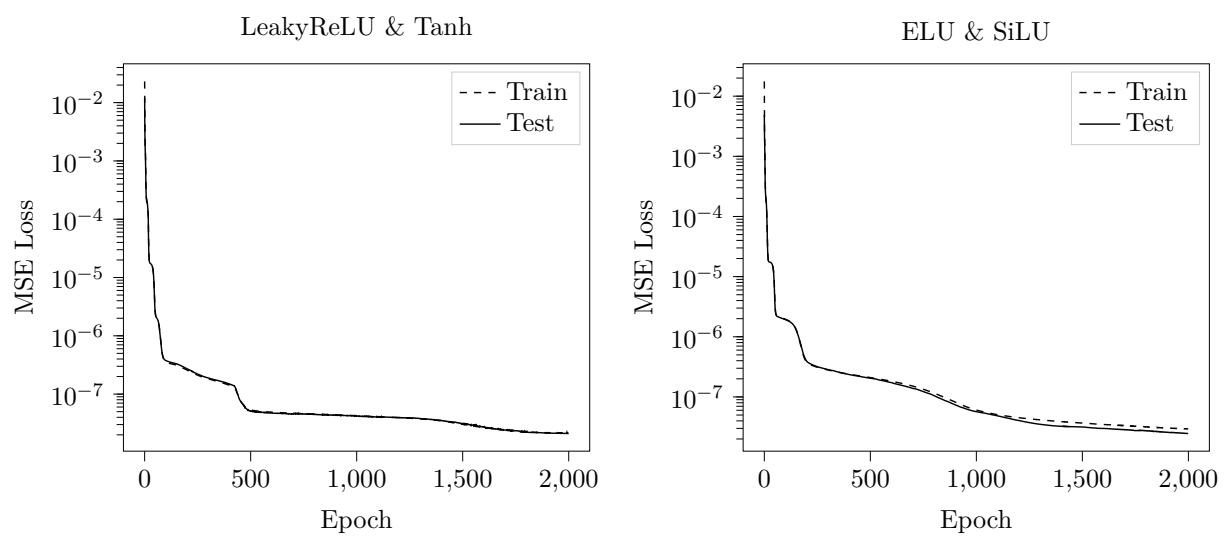


Figure A.5

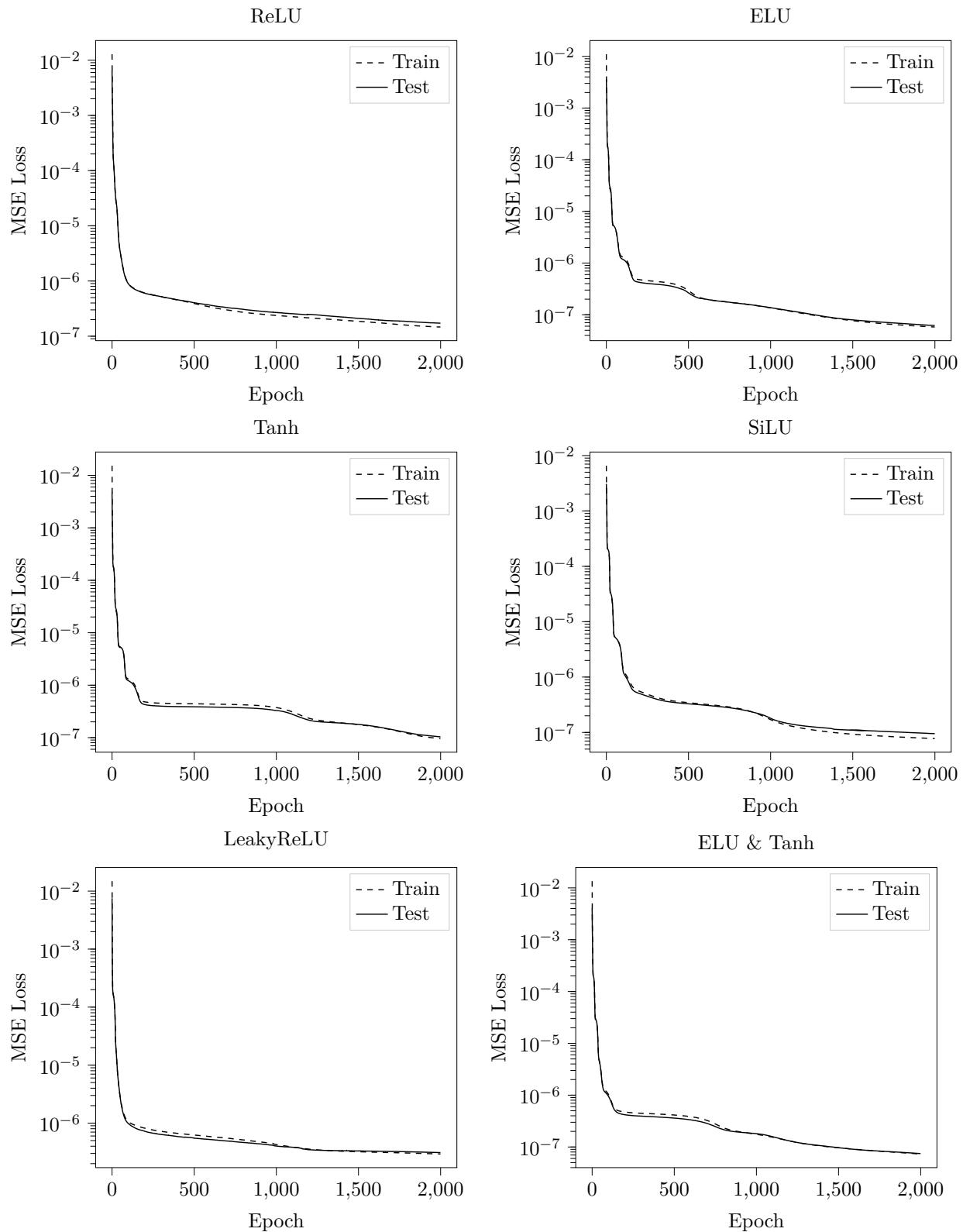


Figure A.6: blablabla

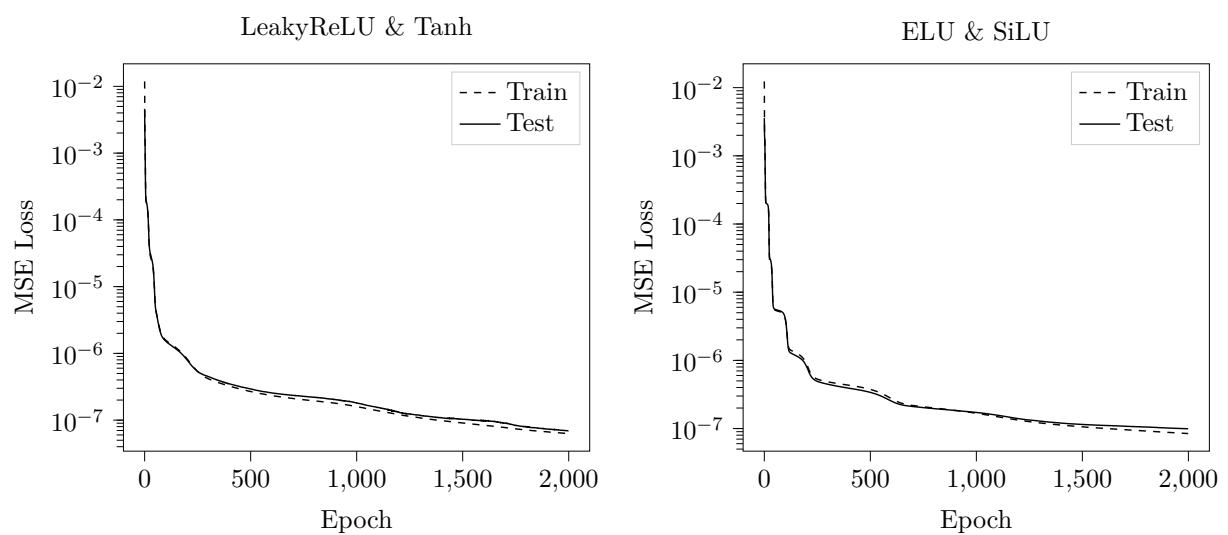


Figure A.6: blablabla

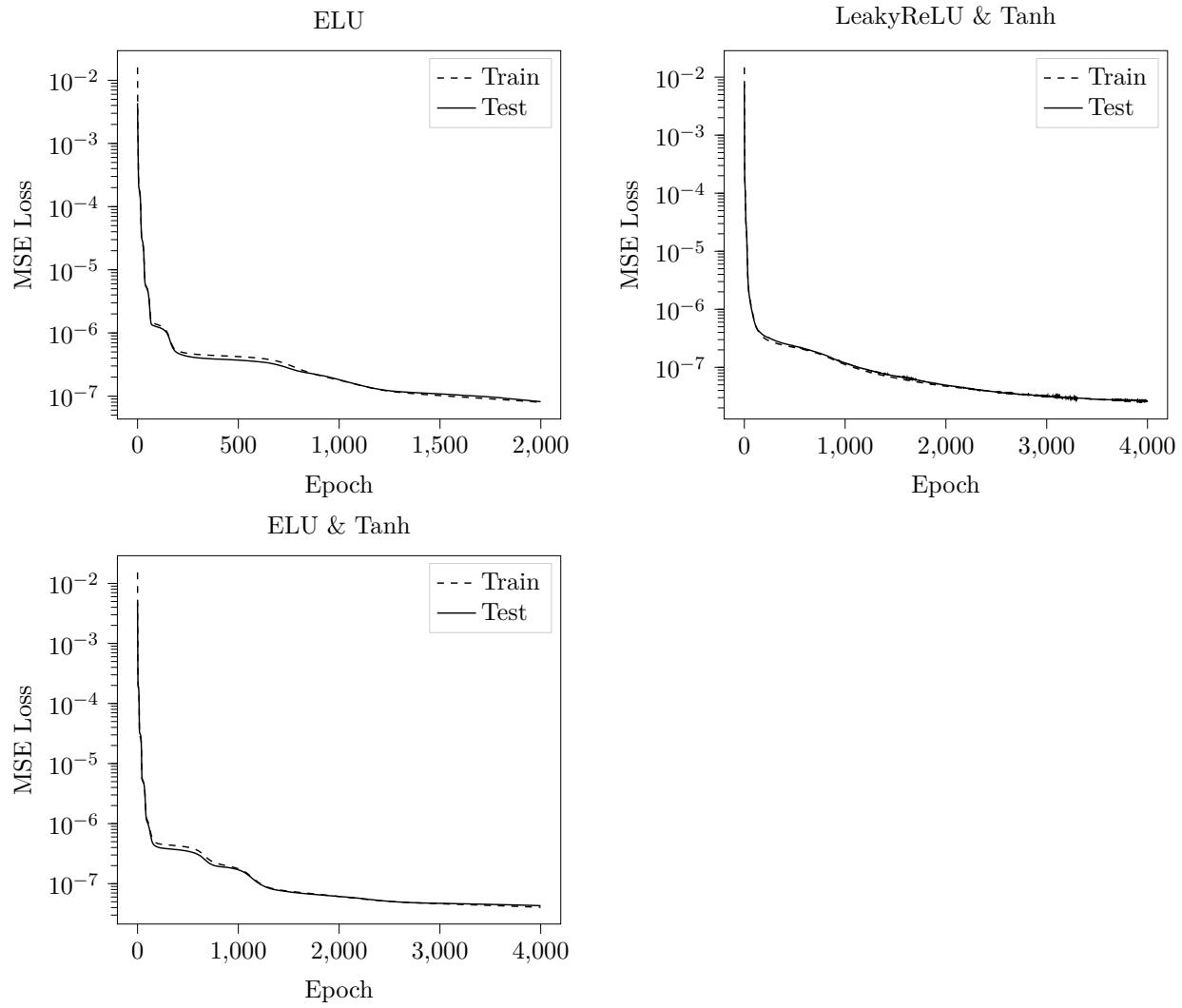
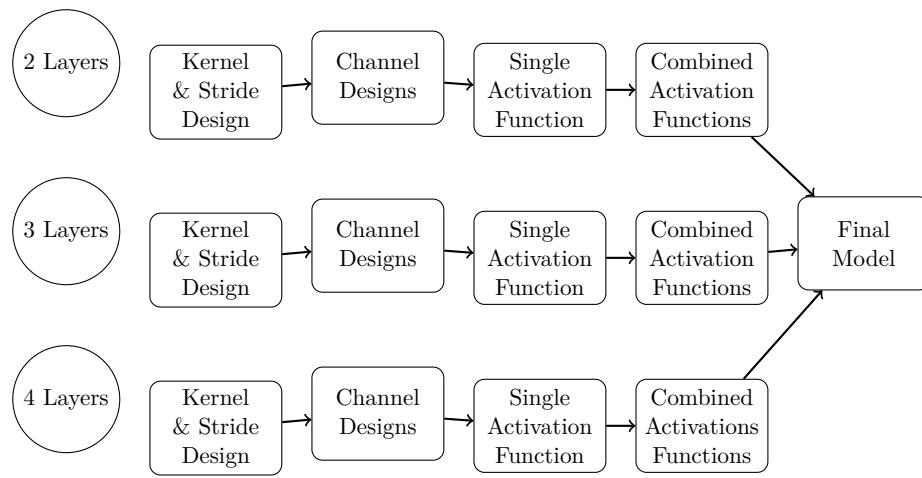


Figure A.7: blablabla

## **B Hyperparameters for the Convolutional Autoencoder**



### **B.0.1 Appendix B**

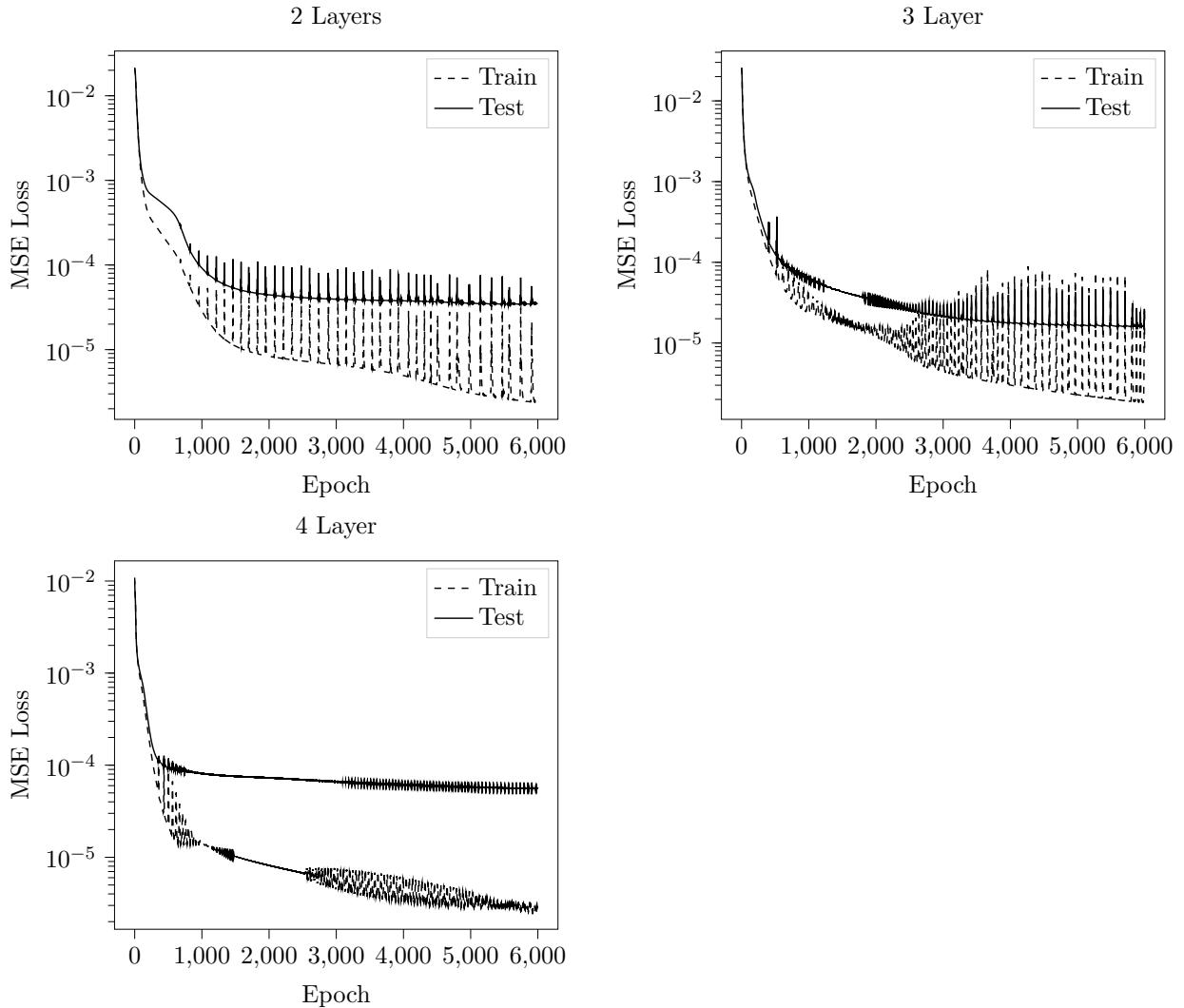
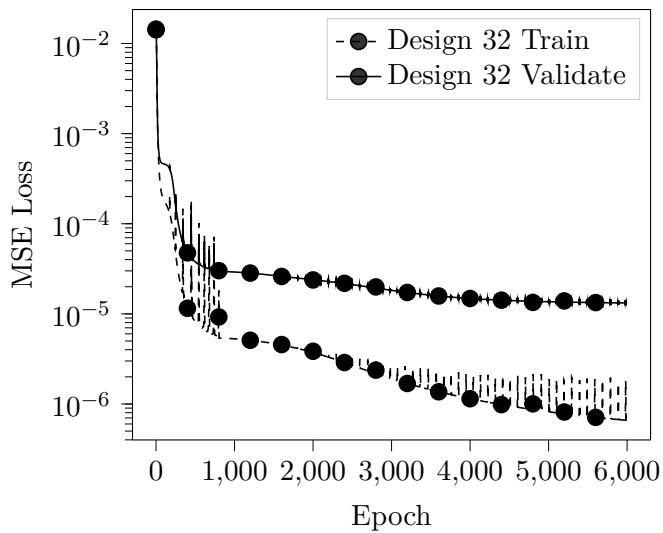
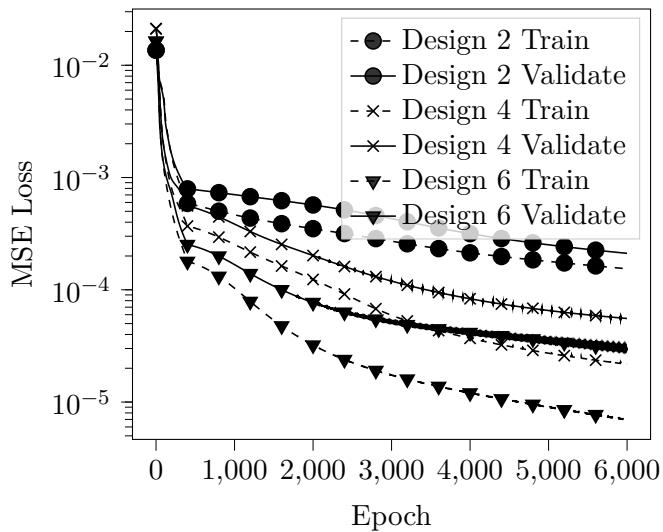


Figure B.1: Three convolutional networks with differing depth and with identical kernel evolution for  $\mathbf{R}$ .

Train & validation for channel designs with 2 layers



Train & validation for channel designs with 3 layers



Train & validation for channel designs with 4 layers

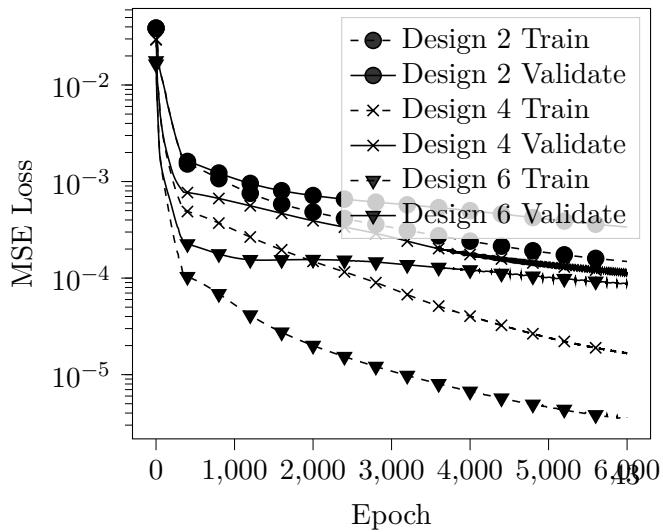


Figure B.2: Different Channel sizes for three convolutional networks with differing depth for  $\mathbf{R}$ .

## Bibliography

---

- [1] Bhatnagar, Gross and Krook, *A model for collision processes in gases*, .
- [2] T. Franz, *Reduced-order modeling of steady transonic flows via manifold learning*. 2016.
- [3] S. L. Brunton and J. N. Kutz, *Data driven science and engineering*. 2019.
- [4] F. Bernard, A. Iollo and S. Riffaud, *Reduced-order model for the bgk equation based on pod and optimal transport*, .
- [5] K. Lee and K. T. Carlberg, *Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders*, .
- [6] I. J. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016.
- [7] D. Rumelhart, G. Hinton and R. Williams, *Learning internal representations by error propagation*, .
- [8] D. H. Ballard, *Modular learning in neural networks*, .
- [9] S. Rifai, P. Vincent, X. Muller, X. Glorot and Y. Bengio, *Contractive auto-encoders: Explicit invariance during feature extraction*, .
- [10] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio and X. Muller, *The Manifold Tangent Classifier*. Curran Associates, Inc., 2011.
- [11] S. Rifai, Y. Bengio, Y. Dauphin and P. Vincent, *A generative process for sampling contractive auto-encoders*, 1206.6434.
- [12] G. Puppo, *Kinetic models of bgk type and their numerical integration*, 1902.08311.
- [13] K. Hornik, M. Stinchcombe and H. White, *Multilayer feedforward networks are universal approximators*, *Neural Networks* **2** (Jan., 1989) 359–366.
- [14] V. Dumoulin and F. Visin, *A guide to convolution arithmetic for deep learning*, 1603.07285.
- [15] M. Ohlberger and S. Rave, *Reduced basis methods: Success, limitations and future challenges*, 1511.02021.
- [16] H. L. Dret and B. Lucquin, *Partial Differential Equations: Modeling, Analysis and Numerical Approximation* -. Birkhäuser, Basel, 2016.

## **Acknowledgement**

## **Hilfsmittel**

## **Selbstständigkeitserklärung**

Ich erkläre, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Seitens des Verfassers bestehen keine Einwände, die vorliegende Masterarbeit für die öffentliche Benutzung im Universitätsarchiv zur Verfügung zu stellen.

Berlin, den 15. Mai 2017

---

Philipp Krah