Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Talgat Daulbaev

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Applications of differential equations and reduced-order modeling for deep learning

Supervisor: Professor Ivan Oseledets
Co-supervisor: Professor Andrzej Cichocki

Name of the Reviewer: Anh-Huy Phan

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**18-03-2023** **Date: DD-MM-YYYY**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

The PhD thesis of Talgat Daulbaev is composed of six chapters, with the organization being significantly improved in this version. The first chapter offers an overview of the problems studied, along with the methodologies and models used, specifically highlighting the Neural Ordinary Differential Equations (Neural ODEs). Chapters 2-6 present the primary contributions that Talgat Daulbaev has proposed and published in five different publications. The dissertation presents a substantial achievement using advanced techniques such as Neural Ordinary Differential Equations and reduced order modeling in deep learning.

1 Major comments

•Chapter 2

The interpolated Reverse Dynamic Method (IRDM) presented in Chapter 2 was proposed to accelerate the training of Neural ODEs. The method has been shown to be more efficient than the standard backpropagation and checkpointing method ANODE in Chapter 2.

Why is IRDM not used in examples in other chapters?

•A key condition for IRDM is that function z(t) has a good approximation by the barycentric Lagrange interpolation. How to check this condition?

The thesis should provide some examples of failure cases.

•What is the relation between memory consumption and stability in computing gradients? What makes the IRDM more stable than RDM?

•The theoretical upper bound on the gradient error is defined in (2.9) and derived in (2.10). How to use this upper bound in practice? When does the upper bound imply a good approximation of the gradients? Examples and experimental results do not provide an analysis of the derived upper bound.

•IRDM needs less number of function evaluations per iteration than RDM, implying that training with IRDM should be faster. The results shown in Figures 2.6, 2.7, and 2.8 do not confirm this advantage of IRDM. The two methods almost demand comparable training time.

•In Chapter 3, page 34

”by more powerful solver we mean ODE solver that requires more right-hand side evaluations to solve”

This statement seems to contradict the motivation and purpose of IRDM.

In Chapter 2, ANODE is criticized because of its instability in computing gradients and a high number of right-hand side evaluations. This motivates the IRDM with a less number of right-hand side evaluations and more stability.

• Chapter 5

What are the typical values of the dimensions Pk? If there exist sketching matrices, $\mathbf{S}_k$ of size $P_k \times D_k$, $P_k \ll D_k$, which can preserve the NN performance, why don’t train a new NN with new smaller kernel sizes $P_k \times P_{k-1}$?

• Since the feature maps are big, instead of using SVD, the max-volume algorithm is used to find the active subspace, V_k, of the layer outputs. However, ranks of V_k are estimated using Variational Bayesian Matrix Factorization, whereas VMBF is based on SVD.

This raises concerns about the efficiency of the proposed method in practice.

• Chapter 6
Contributions presented in Section 6.1.1 are indeed the motivation for the developed methods.

•Error in (6.16), page 68

The matrix comprising r gradients, S of size n × r, r ≪ n, is decomposed by randomized SVD to give $S = U \Sigma V^T$.

After eliminating the last approximated singular vector by the soft-thresholding update in (6.16), the matrix S will have the size of r × (r −1), not n × (r−1).

• Optimization problem in (6.34) and the projection in (6.35)

The optimal solution to the maximization of a convex function over a Euclidean ball must lie on the perimeter of the sphere, i.e., $|| v ||_2 = \delta$. The projection should be $proj_\delta(v) = \delta v / || v ||_2$.

2 Other comments and common errors

•Introduction

The last sentence on page 4 lacks completeness.

”odel reduction to standard artificial neural networks (Chapters 5 and 6).”

- Page 5

”achieve 93% accuracy” -> ”achieve an accuracy of 93%” or ”achieve a

93% accuracy”

”among neural ODEs tested on this problem” -> ”among tested neural

ODEs for this issue”

”arxiv” -> ”arXiv”

”replace convolutional layers with smaller fully-connected layers”

-> ”replace convolutional layers with fully-connected layers of smaller sizes”

- Page 6, page 12 and page 13

”section 1.1”, ”section 1.2”, ”section 1.3” -> ”Section 1.1”, ”Section 1.2”,

”Section 1.3”

problem 1.2 -> problem (1.2)

”chapter 2” -> ”Chapter 2”

”chapter 5” -> ”Chapter 5”

”chapter 6” -> ”Chapter 6”

- Page 9

”if for sufficiently smooth initial value problems problems 1.2 holds”

->

”if the following condition is fulfilled for sufficiently smooth initial-value problems”

- Page 10

”we briefly describe classic approaches to this problem and derive formulas for the so-called adjoint method.”

”classic approaches” ->

”classical approaches” or ”previously established methods”

- Page 11 ”Condition 1.10 ” -> The condition in (1.10)

”Lagrange function 1.8” -> ”Lagrange function in (1.8)”

”by parts one term under the integral”

need to be rephrased

- Page 16

Missing a reference

”We checked this fact by ourselves and got the same results (see ??).”

The figure shown on page 16 has no caption number.

-

”it requires intermediate activations storage and need to perform additional ODE solver steps”

”need” -> ”needs”

- Page 17

”section 1.2” -> ”Section 1.2”

”This method is used in [25], where the neural ODE model is proposed, under the name “adjoint method”.”

- Page 22

”Substitution this equality in (2.9)”

->

”Substituting the above equality into (2.9)”

• Page 23

”smoothness of J(t) (2.19)””

Check the reference for smoothness of J(t).

• Page 24 and page 25

The thesis has no section for supplementary materials.

”the values of optimized hyperparameters are in the supplementary materials”

”In supplementary materials, we provide graphs with empirical results”

• Paragraph ”Improvement in stability of gradient computations”

”To illustrate the stability of the IRDM, we show the plot of test loss vs. training time in density estimation problem, see Figure 2.4a”

The stability of the gradient approximation and interpretation of the obtained results should be analyzed in detail.

• Page 25

”On Figure 2.5” -> In Figure 2.5

• Page 26

”on the training variational autoencoder”

->

”on training variational autoencoder”

”on training of variational autoencoder”

”caltech” -> ”Caltech”

• Page 28, Table 2.1

Are the results shown in the last 6 columns for IRDM?

• Section 2.5.5 for Experimental settings should be moved to the begining

of Section 2.5.

• Chapter 3, page 31

Reference numbers should be enclosed in square brackets.
There is no need to use parentheses for citations.

([25]) -> [25]

([87, 32 ,7]) -> [87, 32 ,7]

• Section 3.1

What is the ODE mentioned in Section 3.1?

”The right-hand side of this ODE”

”to integrate this ODE”

”between steps to integrate this ODE”

• Page 32, about the robustness

Rephrase this part

”batch normalization is often used ..., make it more robust to training hyperparameters”

Does it mean ”BN makes NN robust to the change of hyperparameters”?

• Page 32 and Page 40

”in this study” -> in this chapter”

• Page 33

”are build by stacking”

”built”

”spacial size” -> ”spatial size”

• Page 33

”We use ANODE to train considered models since it is more robust than the adjoint method”

Why is the IRDM not used? Chapter 2 confirms IRDM is faster and more stable than ANODE.

ANODEDEV2 has not been introduced yet in Chapter 3.

• Page 33

”parameters in the right-hand side of the system”

”on the right-hand side”

”freyfaces” -> ”Freyfaces”

•Table 3.1

What is the motivation for the examples using ODENet10 in Section 3.3 and Table 3.1 if the performance accuracy is not better than ResNet10?

•
”the right-hand side evaluations necessary for integration of system (2.1),”

the system in (3.1)

• Figure 3.1

Discussion on the results in Figure 3.1 should be moved to the main text.

•Page 37

”As we already mentions”

•Page 38

The Butcher table has no caption and table number.

•Page 43

”Table 4.2, 4.4”

”Tables 4.2 and 4.4”

•Page 55,

Remove the redundant sub-figure number ”[b]” in Figure 5.2.

•Page 55

”approach at 0.3For the experiments”

Put a full stop before ”For”

• Page 62

in (6.2), use ”. . . ” in place of ”
. . .
”
•Definition of the active subspace is not very accurate

”The subspace spanned by matrix $V_1 \in Rn \times r$ is called an active subspace because c(x) is sensitive to perturbation vectors within this subspace”

It is obvious that $[V_1, v_{r+1}]$ also gives another active subspace, unless $\lambda_{r+1} = 0$.

•Page 66
”a small active neurons”

->

”a small number of active neurons”

•Page 75 , Algorithm 4

”dv” -> ”d_v”

•Reference list

Remove duplicate references, [21], [22], [107], [108], [150], [151], [156], [157]

Provisional Recommendation

☒ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate’s thesis according to the recommendations of the present report*

☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*