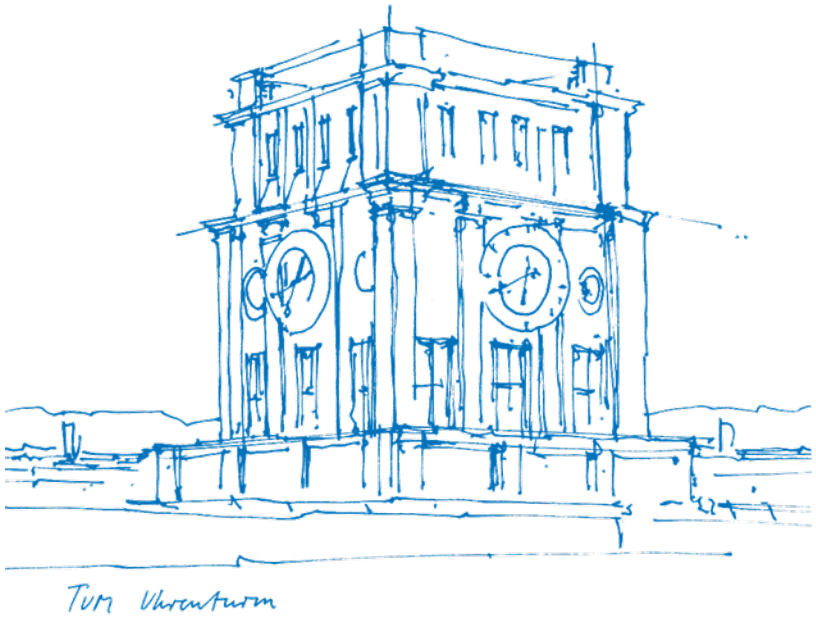
Machine Learning - looking ahead

Dr. Felix Dietrich

2020-01-09



Organizational issues

Groups, Moodle, Reports

- Did every group upload their reports for the exercise?
- Let me know if you have questions about my feedback.
- You will receive the points for the exercise in the next two weeks.

Highlights

Exercise 4

- A lot of trouble with Diffusion Maps, with many small bugs
- Two groups used the sklearn interface for their code - very good![1]
- Many great visualizations
- An example implementation for the VAE is online on Moodle, take a look!

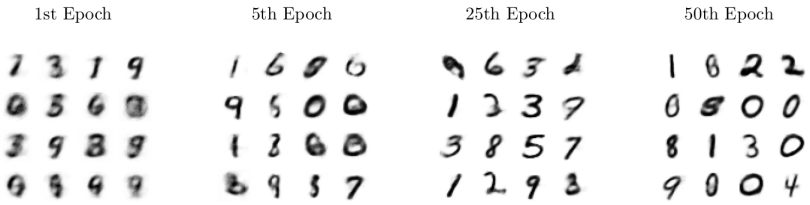
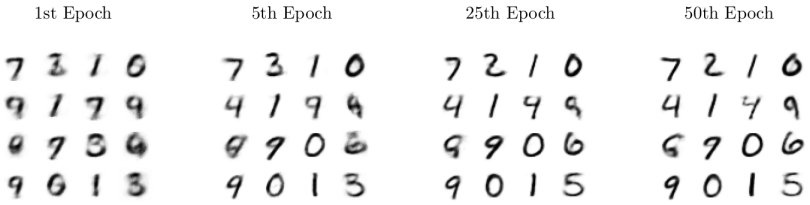


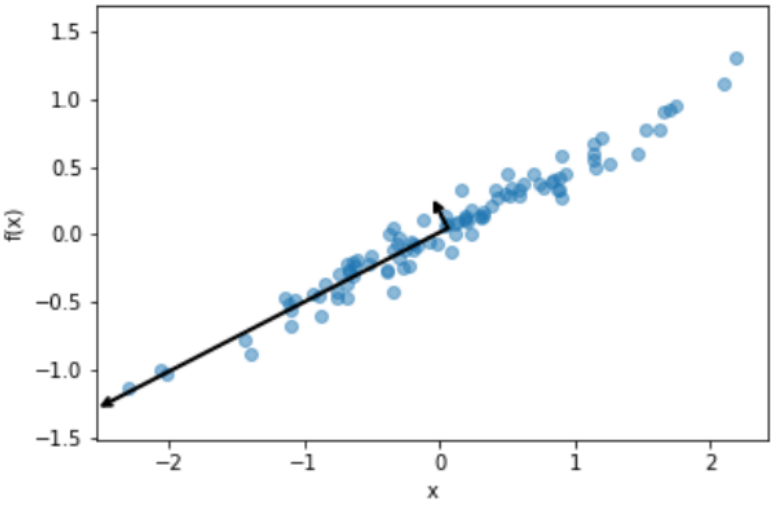
Figure 11: Reconstructed results obtained with a 32 dimensional latent space



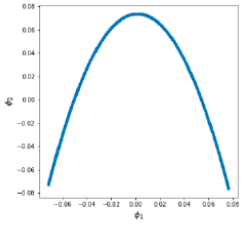
[1] Take a look: `https://github.com/scikit-learn/scikit-learn/blob/e5698bde9/sklearn/decomposition/_pca.py#L104`

Highlights

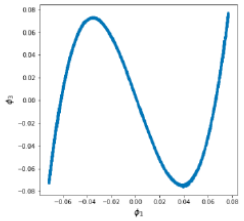
Exercise 4



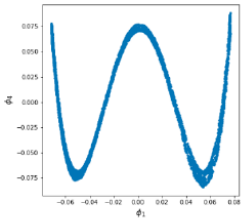
(b) Direction of 2 PC marked in original data-set.



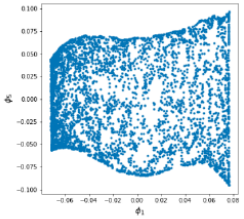
(a) ϕ_1 against ϕ_2



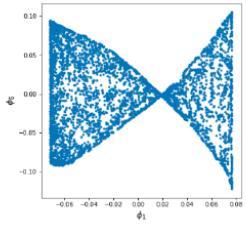
(b) ϕ_1 against ϕ_3



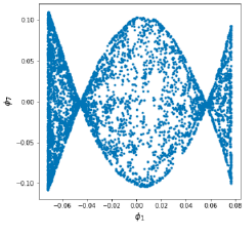
(c) ϕ_1 against ϕ_4



(d) ϕ_1 against ϕ_5



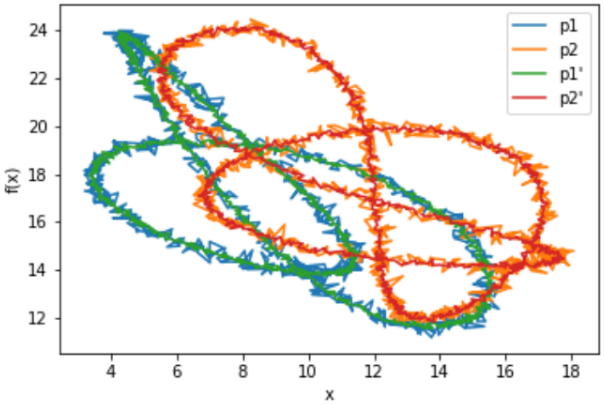
(e) ϕ_1 against ϕ_6



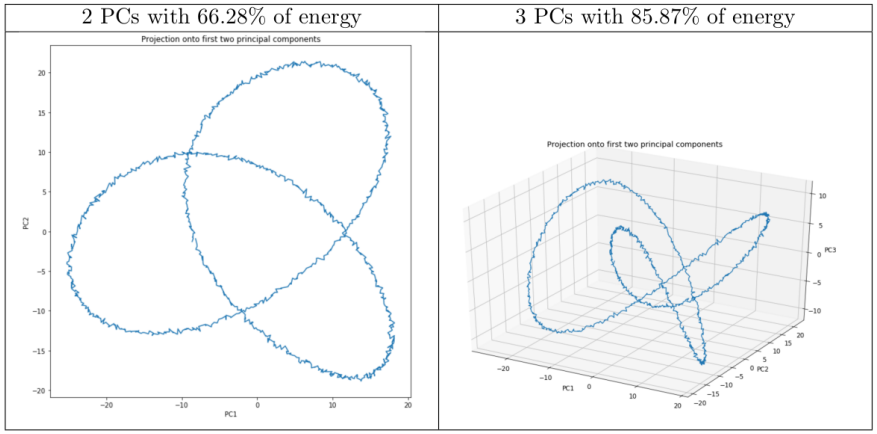
(f) ϕ_1 against ϕ_7

Highlights

Exercise 4



(b) Original dataset with projected 4 PC on the dataset.



Recap of the course

Lecture 1: Modeling crowd dynamics

- Modeling approaches, verification and validation

Lecture 2: Simulation software

- Introduction to the Vadere software (guest lecture from Benedikt Zönnchen)

Lecture 3: Bifurcation theory and visualization

- Dynamical systems and bifurcation theory

Lecture 4: Representation of data

- Principal Component Analysis, Diffusion Maps, VAE (guest lecture from Alexej Klushyn)

Lecture 5: Extracting dynamical systems from data

- Function approximation, vector fields, time-delay embedding

Today: Machine Learning, looking ahead

Future directions

1. Challenges in data science
2. Jobs in data science
3. Master's thesis topics
4. Final projects

Today: Machine Learning, looking ahead

Challenges in data science

1. Integrating knowledge, physics into algorithms
2. Using context of the data
3. Extracting knowledge understandable for humans, visualization
4. Managing and handling heterogenous data
5. Handling (lack of) real data - the GO software cannot be used in reality
6. Reducing the carbon footprint
7. Moving ML from computer engineering to computer science and mathematics

Challenges in data science

Integrating domain knowledge into algorithms

State of the art:

- Supervised learning (function approximation)
- Unsupervised learning (manifold, distribution learning)
- Generative models (models of random variables)

Challenges in data science

Integrating domain knowledge into algorithms

State of the art:

- Supervised learning (function approximation)
- Unsupervised learning (manifold, distribution learning)
- Generative models (models of random variables)

Goals:

1. Algorithms adapted to context
2. Physics/biology/chemistry/... “built-in”

Challenges in data science

Integrating domain knowledge into algorithms

State of the art:

- Supervised learning (function approximation)
- Unsupervised learning (manifold, distribution learning)
- Generative models (models of random variables)

Goals:

1. Algorithms adapted to context
2. Physics/biology/chemistry/... “built-in”

Resources:

1. The Science of Deep Learning `https://www.youtube.com/playlist?list=PLGJm1x3XQeK0gmqfRkP-VmrEf4UYx5IDW`
2. Using Physical Insights for Machine Learning `https://www.ipam.ucla.edu/programs/workshops/workshop-iv-using-physical-insights-for-machine-learning/?tab=schedule`
3. An example: “On learning Hamiltonian systems from data” [Bertalan et al., 2019]

Challenges in data science

Managing heterogeneous data

State of the art:

- Scattered data sources
- Heterogeneous file formats
- Proprietary codes

Challenges in data science

Managing heterogeneous data

State of the art:

- Scattered data sources
- Heterogeneous file formats
- Proprietary codes

Goals: standards, open data, global availability

Challenges in data science

Managing heterogeneous data

State of the art:

- Scattered data sources
- Heterogeneous file formats
- Proprietary codes

Goals: standards, open data, global availability

Resources:

1. Pedestrian counting system in Melbourne `http://www.pedestrian.melbourne.vic.gov.au`
2. GBIF—the Global Biodiversity Information Facility `https://www.gbif.org/dataset/search`
3. The Novel Materials Discovery (NOMAD) Laboratory `https://nomad-coe.eu/`
4. Nomad2018 Predicting Transparent Conductors on kaggle `https://www.kaggle.com/c/nomad2018-predict-transparent-conductors`
5. IPAM talk about LHC challenges `http://www.ipam.ucla.edu/abstract/?tid=15030&pcode=BDCWS2`

Machine Learning, looking ahead

Jobs: Data Scientist

In general: highly interactive, many different disciplines work together

Industry:

- Many companies need ML
- Competition against people who work in ML for 10+ years
- Microsoft, Google, Facebook, ... consider TUM graduates a lot
- Mobility not that important

Machine Learning, looking ahead

Jobs: Data Scientist

In general: highly interactive, many different disciplines work together

Industry:

- Many companies need ML
- Competition against people who work in ML for 10+ years
- Microsoft, Google, Facebook, ... consider TUM graduates a lot
- Mobility not that important

Academia:

- Broad range of fields need ML
- Relatively easy to start in applied sciences as computer scientist/mathematician
- Mobility a requirement
- Ask me about career paths!

Machine Learning, looking ahead

Master's thesis topics

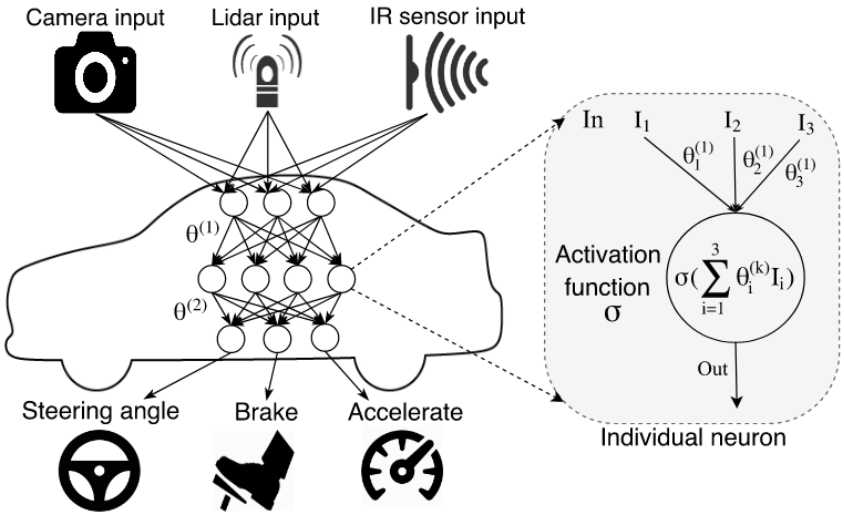
1. Model predictive control: autonomous cars
2. Model predictive control: landing a rocket
3. Image recognition and classification
4. High performance computing in Machine Learning
5. Gaussian process landmarks
6. Optimization with surrogate models
7. Context-aware ML: Using context to unwrap paper and study bacteria
8. Neural networks solving linear systems
9. Hazard identification in heterogeneous sensors
10. ... (topics in crowd dynamics, etc.)

Master's thesis

Topic 1: Model predictive control: autonomous cars

1. Understand MPC
2. Understand car test environment
3. Setup the test environment
4. Setup MPC
5. Steer a (simulated) car autonomously!

References: `http://carla.org/`;
`http://apollo.auto/platform/simulation.html`;
`https://github.com/tawnkramer/gym-donkeycar`;
[Tian et al., 2018]



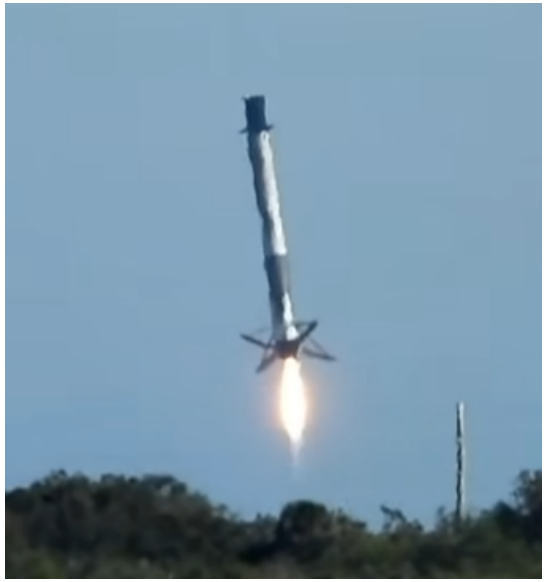
From [Tian et al., 2018], figure 2.

Master's thesis

Topic 2: Model predictive control: landing a rocket

1. Understand MPC
2. Understand test environment "Kerbal space program"
3. Setup the test environment
4. Setup MPC
5. Land a (simulated) rocket!

References: [Lovelett et al., 2018];
`https://www.kerbalspaceprogram.com/`;
`https://www.youtube.com/watch?v=wNpDWSfSd8k&feature=youtu.be&t=160`



From `https://www.youtube.com/watch?v=wNpDWSfSd8k&feature=youtu.be&t=160`

Master's thesis

Topic 3: Image recognition and classification

1. Understand convolution in neural networks
2. Understand Gaussian process regression
3. Implement convolutional Gaussian process regression
4. Image classification using Gaussian process regression

References: [Wang and Raj, 2017, Sindagi and Patel, 2017, Cohen et al., 2018, Gao et al., 2019]

From `https://www.publicdomainpictures.net/en/view-image.php?image=123418&picture=7-cats-in-various-colours`

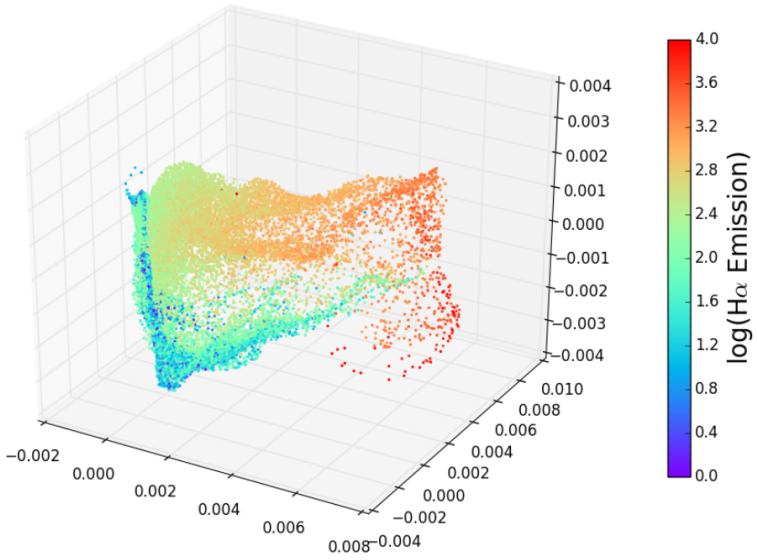
Master's thesis

Topic 4: High performance computing in Machine Learning

1. Read the megaman paper [McQueen et al., 2016]
2. Reproduce their results with their software
3. Compare results against SpectralNet, Diffusion Variational Autoencoders
4. (Research) Analysis of algorithmic complexity for all approaches
5. (Research) Compare to [Pflüger et al., 2010], "Spatially adaptive sparse grids for high-dimensional data-driven problems"

References:

[Pflüger et al., 2010, McQueen et al., 2016, Shaham et al., 2018, Li et al., 2019]



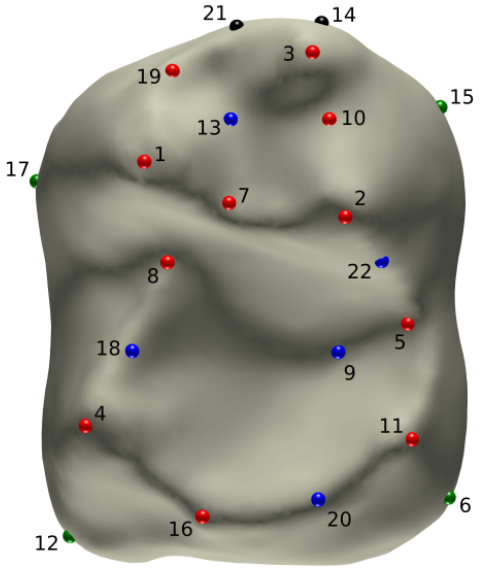
From [McQueen et al., 2016], figure 6.

Master's thesis

Topic 5: Gaussian process landmarks

1. Understand Gaussian process regression
2. Understand the concept of landmarks
3. Reproduce the results in [Gao et al., 2019]
4. Suggest other approaches and implement them
5. (More applied) Use in image classification
6. (More theoretical) Compare to [Dunson et al., 2019], "Diffusion based Gaussian process regression"

References:
[Gao et al., 2019, Dunson et al., 2019]



(b) Gaussian Process Landmarks

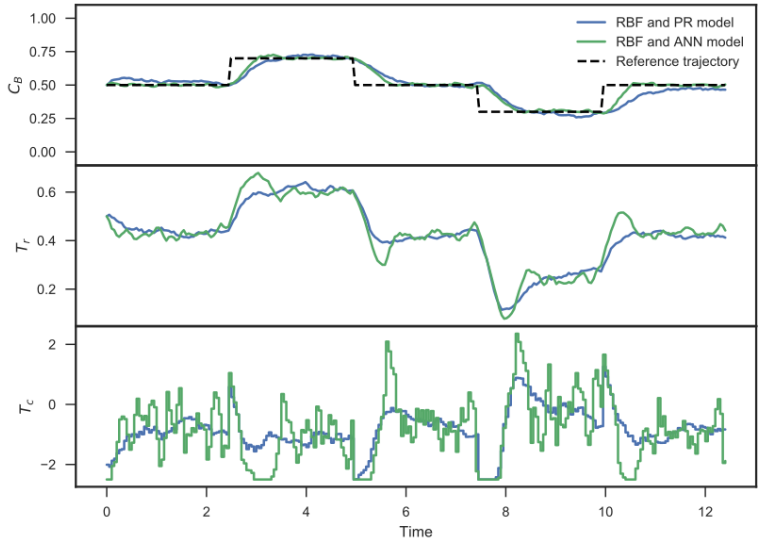
From [Gao et al., 2019], figure 2.

Master's thesis

Topic 6: Optimization with surrogate models

1. Understand Model Predictive Control
2. Understand Optimal Control
3. Implement MPC with a surrogate model
4. Implement Optimal control with a surrogate model
5. (More applied) see autonomous cars, rockets
6. (Optional) Understand / implement Koopman operator framework + control

References: [Lovelett et al., 2018, Korda and Mezić, 2018, Kaiser et al., 2017]



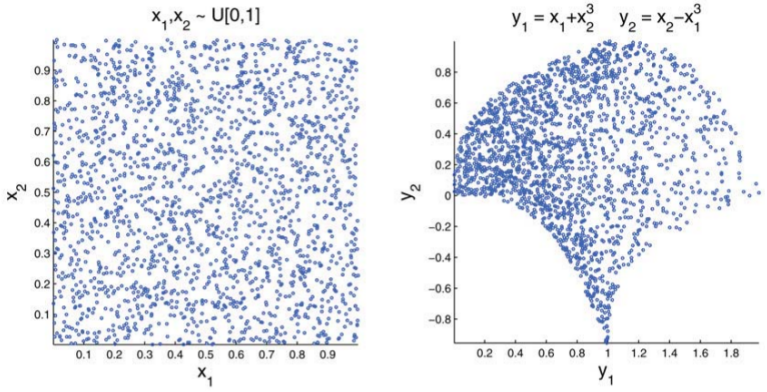
From [Lovelett et al., 2018], figure 6.

Master's thesis

Topic 7: Context-aware ML

1. Understand nonlinear independent component analysis
2. Compare to traditional independent component analysis
3. Implement the methods and apply them to toy examples
4. (Research oriented) work with me on unpublished results
5. (More applied) Unwrap paper from image data
6. (More applied) Study bacteria from fluorescence data

References:
[Singer and Coifman, 2008, Singer et al., 2009] + unpublished



From [Singer and Coifman, 2008], figure 1.

Master's thesis

Topic 8: Neural networks solving linear systems

1. Understand "Neural networks for solving systems of linear equations and related problems"
2. Implement a neural network that solves linear systems
3. Compare results to standard linear solvers
4. (More theoretical) Analysis of algorithmic complexity
5. (More applied) Numerical comparison of algorithmic complexity
6. (Research) Analyze SpectralNet and Diffusion Autoencoder

References: [Cichocki and Unbehauen, 1992, Shaham et al., 2018, Li et al., 2019]

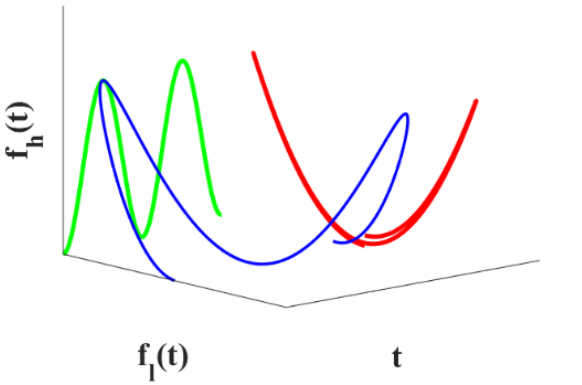
From [Li et al., 2019], figure 3.

Master's thesis

Topic 9: Hazard identification in heterogeneous sensors

1. Understand multi-fidelity modeling and sensor data fusion
2. Reproduce toy example results from [Lee et al., 2019]
3. Apply data fusion to heterogeneous data from crowds (real and simulated)
4. Use multi-fidelity modeling to identify hazards for crowds
5. (Research) Compare data fusion [Williams et al., 2015] with multi-fidelity modeling

References: [E et al., 2007, Williams et al., 2015, Perdikaris et al., 2017, Lee et al., 2019]



(a)A correlation between t, $f_l(t)$, and $f_h(t)$

From [Lee et al., 2019], figure 2.

Final projects

Presentations - general

1. 20 min presentation + 5 min questions
2. Everyone in every group has to present
3. You have to upload your report on Moodle (as usual) before your presentation
4. Grading: 1/3 report, 1/3 code, 1/3 presentation. 25% of total grade!

Final projects

Presentations - general

1. 20 min presentation + 5 min questions
2. Everyone in every group has to present
3. You have to upload your report on Moodle (as usual) before your presentation
4. Grading: 1/3 report, 1/3 code, 1/3 presentation. 25% of total grade!

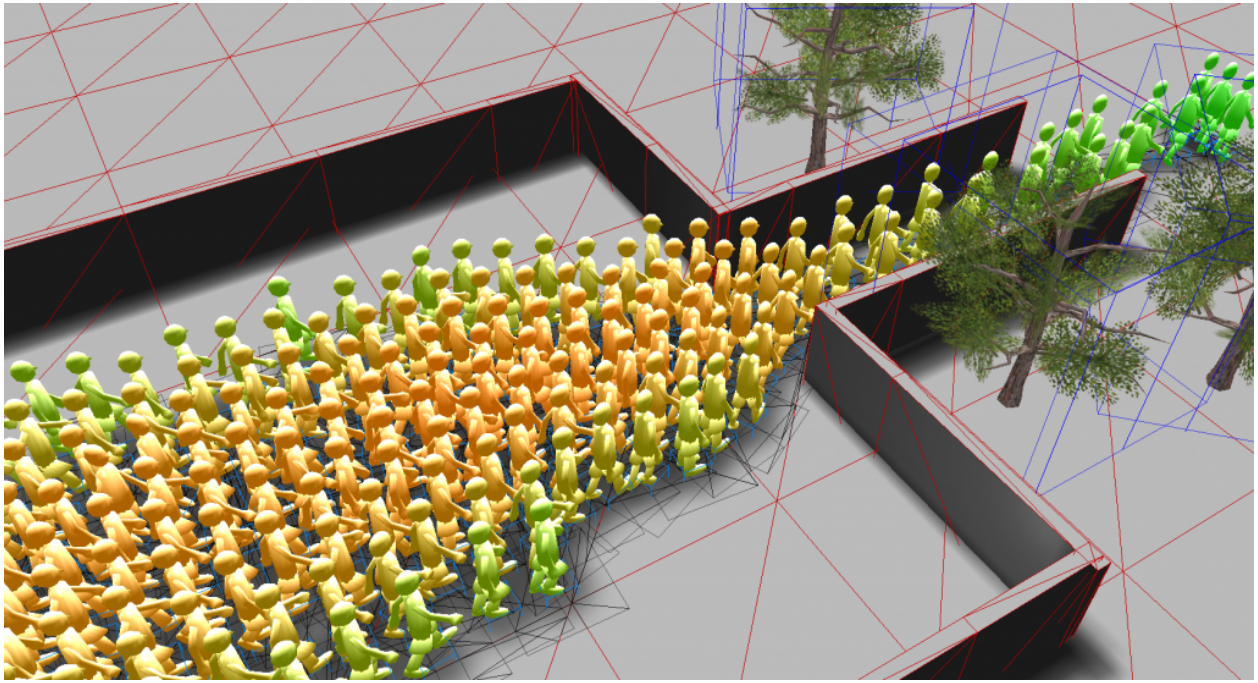
Presentations - 2020-01-23 (work about the same as the usual exercise)

1. Group B: Learning dynamical systems from data: Neural networks
2. Group J: Learning dynamical systems from data: Neural networks
3. Group F: Review and implementation of “Neural Ordinary Differential Equations”
4. Group C: An efficient, robust and functional visualization for crowd trajectories

Presentations - 2020-01-30 (I expect you do 50% more than the usual exercise)

1. Group H: Implementing and testing the SocialGAN network
2. Group E: Learning dynamical systems from data: Koopman operator
3. Group G: Review and implementation of “megaman: Manifold learning with millions of points”
4. Group D: Review and implementation of “megaman: Manifold learning with millions of points”

Questions?



Homework: work on your final group project and present at the date we agreed on.
Homework (optional): think about a Master's thesis in data science.
For questions / appointments: please ask via email, `felix.dietrich@tum.de`.

Literature I

Bertalan, T., Dietrich, F., Mezić, I., and Kevrekidis, I. G. (2019).
On learning hamiltonian systems from data.
Chaos: An Interdisciplinary Journal of Nonlinear Science, 29(12):121107.

Cichocki, A. and Unbehauen, R. (1992).
Neural networks for solving systems of linear equations and related problems.
IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications.

Cohen, T. S., Geiger, M., Koehler, J., and Welling, M. (2018).
Spherical cnns.
In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Dunson, D. B., Wu, H.-T., and Wu, N. (2019).
Diffusion based gaussian process regression via heat kernel reconstruction.
arXiv.

E, W., Engquist, B., Li, X., Ren, W., and Vanden-Eijnden, E. (2007).
Heterogeneous multiscale methods: A review.
Communications in Computational Physics.

Gao, T., Kovalsky, S. Z., and Daubechies, I. (2019).
Gaussian process landmarking on manifolds.
SIAM Journal on Mathematics of Data Science, 1(1):208–236.

Kaiser, E., Kutz, J. N., and Brunton, S. L. (2017).
Data-driven discovery of koopman eigenfunctions for control.
arXiv.

Korda, M. and Mezić, I. (2018).
Optimal construction of koopman eigenfunctions for prediction and control.
arXiv.

Lee, S., Dietrich, F., Karniadakis, G. E., and Kevrekidis, I. G. (2019).
Linking gaussian process regression with data-driven manifold embeddings for nonlinear data fusion.
Interface Focus, 9(3):20180083.

Li, H., Lindenbaum, O., Cheng, X., and Cloninger, A. (2019).
Diffusion variational autoencoders.
arXiv.

Lovelett, R. J., Dietrich, F., Lee, S., and Kevrekidis, I. G. (2018).
Some manifold learning considerations towards explicit model predictive control.
arXiv.

Literature II


McQueen, J., Meila, M., VanderPlas, J., and Zhang, Z. (2016).
megaman: Manifold learning with millions of points.
arXiv.


Perdikaris, P., Raissi, M., Damianou, A., Lawrence, N., and Karniadakis, G. (2017).
Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling.
Proceedings Royal Society A, 473(2198).


Pflüger, D., Peherstorfer, B., and Bungartz, H.-J. (2010).
Spatially adaptive sparse grids for high-dimensional data-driven problems.
Journal of Complexity.


Shaham, U., Stanton, K., Li, H., Nadler, B., Basri, R., and Kluger, Y. (2018).
Spectralnet: Spectral clustering using deep neural networks.
arXiv.


Sindagi, V. A. and Patel, V. M. (2017).
A survey of recent advances in cnn-based single image crowd counting and density estimation.
Pattern Recognition Letters.


Singer, A. and Coifman, R. R. (2008).
Non-linear independent component analysis with diffusion maps.
Applied and Computational Harmonic Analysis, 25(2):226–239.


Singer, A., Erban, R., Kevrekidis, I. G., and Coifman, R. R. (2009).
Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps.
Proceedings of the National Academy of Sciences, 106:16090–16095.


Tian, Y., Pei, K., Jana, S., and Ray, B. (2018).
Deeptest: automated testing of deep-neural-network-driven autonomous cars.
ICSE '18: Proceedings of the 40th International Conference on Software Engineering.


Wang, H. and Raj, B. (2017).
On the origin of deep learning.
arXiv.


Williams, M. O., Rowley, C. W., Mezić, I., and Kevrekidis, I. G. (2015).
Data fusion via intrinsic dynamic variables: An application of data-driven koopman spectral analysis.
EPL (Europhysics Letters), 109(4).