Topics in the Foundations of Artificial Intelligence

A Reader

Draft from April 17, 2024.
Please don't cite or distribute without permission.
Comments welcome!

Levin Hornischer
Levin.Hornischer@lmu.de
https://levinhornischer.github.io/FoundAI/

Contents

Preface

This is the reader for the course *Advanced Topics in the Foundations of AI* given during the summer semester 2024 at *LMU Munich* as part of the *Master in Logic and Philosophy of Science*. The reader is written as the course progresses. A website for the course is found at

`https://levinhornischer.github.io/FoundAI/`.

Comments I'm happy about any comments: spotting typos, finding mistakes, pointing out confusing parts, or simply questions triggered by the material. Just send an informal email to `Levin.Hornischer@lmu.de`.

Content In recent years, artificial intelligence and, in particular, machine learning made great—but also disconcerting—progress. However, their foundations are, unlike other areas of computer science, less well understood. This situation is sometimes compared to being able to build steam engines without having a theory of thermodynamics.

This seminar is about the mathematical foundation of AI. After a review of the classical theory (Computability Theory, No-Free-Lunch Theorem, Universal Approximation Theorem, etc.), we read some recent research papers to get an overview of some current approaches to the foundations of AI.

Objectives In terms of content, the course aims to convey and overview of the foundations of AI—including both classic material and cutting-edge research. In terms of skills, the course aims to teach the ability to both mathematically and philosophically assess the different approaches to the foundations of AI.

Prerequisites In order to appreciate the literature, the course requires basic familiarity with mathematics (calculus, linear algebra, probability theory), logic (including, ideally, computability theory), and AI (neural networks). Some papers also use more advanced concepts from topology, probability theory, or category theory, so you should also be prepared to read up on those. But they are not assumed: the seminar sessions are,

among others, meant to get clearer on these concepts. Programming skills will of course be useful, but will not be assumed.

Schedule and organization The course is organized as a seminar. Hence, for each session, we have assigned readings, which we then discuss during the session. The reading for each week is announced in the schedule on the course's website. The readings are roughly organized by topic, forming the chapters of this reader.

Background material Some helpful short explainer videos on AI are found here. A excellent mini series on (the mathematics of) neural networks is found here.

Moreover, the material of my companion course on the Philosophy and Theory of AI might also be helpful. You can take the present course independently of that companion course and vice versa, but they do complement each other. The companion course is more introductory and looks at a broader range of philosophical issues connected to AI and how to theorize about them, while the present course focuses specifically on the more mathematical foundations of neural networks.

A recent edited collection on the mathematical foundations of AI is Grohs and Kutyniok (2022).

Layout These notes are informal and partially still under construction. For example, there are margin notes to convey more casual comments that you'd rather find in a lecture but usually not in a book. Todo notes indicate, well, that something needs to be done. References are found at the end.

This is a margin note.

This is a todo note

Notation Throughout, 'iff' abbreviates 'if and only if'.

1 Introduction

Summary

We give an outlook of the course and of this document.

The field of AI is typically characterized along the lines of aiming to build "machines that can compute how to act effectively and safely in a wide variety of novel situations" (Russell and Norvig 2021, p. 19). In chapter 2, we briefly collect basic terminology and concepts in AI, to make sure we're all on the same page.

Regardless of the definition, it is helpful to distinguish two main traditions in AI. They go by varying names, with different connotations depending on the community that uses them, including the following.

1. *Symbolic AI*: classicist, logic-based, Good Old-Fashioned AI (GOFAI), etc.

   Example: An algorithm or computer program that, given as input a position in a game of chess, outputs the next best move. This algorithm was written by a programmer.

2. *Subsymbolic AI*: connectionist, non-logicist, machine learning, deep learning, etc.

   Example: A neural network that, given as input a pixel image of a handwritten digit, outputs the digit depicted on the image. The neural network was trained to become better at this mapping using thousands of data points, i.e., input images labeled with the digit depicted on them.

(Some might distinguish a third tradition—*statistical AI*—which, in a sense, sits between the two preceding traditions: Like symbolic AI it typically has 'interpretable' variables, but they are now continuous random variables, and like subsymbolic AI it typically processes the information in a continuous way.)

For symbolic AI, we have a pretty good theory due to mathematical/philosophical logic and computability theory. We review this in chapter 3. It gives a good idea of what we also would expect of a theory of AI.

For subsymbolic AI—which is most of modern AI—we, however, lack a good theory. So in this course, we mostly focus on approaches to provide such a theory. In chapter 4, we review the main results that the standard theory of machine learning—which is mostly statistical learning theory—can deliver. But we also look at what is still missing for the concrete case of the neural networks that modern AI is built on.

The remaining chapters—as listed in the table of contents—then are about different approaches to fill these gaps in the standard theory of machine learning, or approaches that rethink this theory all together. There is more material than we will be able to cover: especially from the later chapters we will pick the topics based on your interests.

2 Background

Readings

- A textbook introduction to the field of AI: Russell and Norvig (2021, ch. 1).

Key concepts

- History of AI (summers and winters)
- Types of AI: symbolic, subsymbolic, statistical
- Definitions of AI: acting humanly (Turing test), thinking humanly, acting rationally, thinking rationally
- Types of learning tasks: Supervised learning, unsupervised learning, reinforcement learning. Machine learning pipeline (conceptualization, data, model, deployment).
- Key concepts of artificial neural networks: neurons, layers, feedforward/recurrent, weights, activation function, loss function, backpropagation, learning rate, local/global minima (equilibrium), regularization, overfitting/underfitting.

In this session, we discuss the main reading to get an understanding of each of the key concepts—the basic AI terminology—mentioned above. These key concepts are further illuminated in the additional material mentioned below.

Further material

- A very accessible overview, written at the beginning of the deep learning revolution: Boden (2016, ch. 1 and 4).
- The Stanford Encyclopedia of Philosophy entry on artificial intelligence: Bringsjord and Govindarajulu (2024).

- Also see the background material mentioned in the preface (explainer videos and companion course).

3 Foundations of symbolic AI

Readings

- An overview of symbolic AI: Flasiński (2016, ch. 2)
- An overview of computability and complexity theory: Immerman (2021)

Key concepts

- TBA

Further material

- TBA

4 Standard theory of machine learning

Readings

- Shalev-Shwartz and Ben-David (2014). Chapter 5 on the No-Free-Lunch Theorem and chapter 6 on the fundamental theorem of statistical learning theory.
- One of the classic sources on the universal approximation theorem: Hornik et al. (1989).

Key concepts

- Statistical learning theory
- Universal approximation theorem
- No-Free-Lunch theorem

Further material

- A modern, more general approach to the universal approximation theorem: Kratsios (2021).
- A discussion of the no-free-lunch theorem: Sterkenburg and Grünwald (2021).

5 Refining statistical learning theory

Readings

- Berner et al. (2022)
- Belkin (2021)

Key concepts

- TBA

Further material

- TBA

6 Computability theory of machine learning

Readings

- Colbrook et al. (2022)
- Caro (2023)

Key concepts

- TBA

Further material

- An older overview: Šíma and Orponen (2003).
- Delétang et al. (2023)

7 Using statistical mechanics

Readings

- Roberts and Yaida (2022)
- Bahri et al. (2020)

Key concepts

- TBA

Further material

- The book by Roberts and Yaida (2022) has been presented in a course (`https://deeplearningtheory.com/lectures/`).

8 Topological data analysis

Readings

- Naitzat et al. (2020)
- Overview: Hensel et al. (2021)

Key concepts

- TBA

Further material

- For a short explanation of persistent homology, see here.
- For a popular science application to neuroscience, see here.

9 Category theory as a language of machine learning

Readings

- Shiebler et al. (2021)
- Bradley et al. (2021)
- van Bekkum et al. (2021)

Key concepts

- TBA

Further material

- More on categories for AI in this course (`https://cats.for.ai/`) and this list of papers on the topic (`https://github.com/bgavran/Category_Theory_Machine_Learning`).

10 Dynamical systems

Readings

- Overview: Bournez and Pouly (2021)

Key concepts

- TBA

Further material

- TBA

11 Verification of neural networks

Readings

- Albarghouthi (2021)

Key concepts

- TBA

Further material

- TBA

12 Post-hoc explainability

Readings

- Bilodeau et al. (2024)

Key concepts

- TBA

Further material

- TBA

Bibliography

Albarghouthi, A. (2021). *Introduction to Neural Network Verification*. arXiv: 2109.10317 [cs.LG] (cit. on p. 15).

Bahri, Y. et al. (2020). "Statistical Mechanics of Deep Learning." In: *Annual review of condensed matter physics* 11.1, pp. 501–528. DOI: 10.1146/annurev-conmatphys-031119-050745 (cit. on p. 11).

Belkin, M. (2021). "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation." In: *Acta Numerica* 30, pp. 203–248 (cit. on p. 9).

Berner, J. et al. (2022). "The Modern Mathematics of Deep Learning." In: *Mathematical Aspects of Deep Learning*. Ed. by P. Grohs and G. Kutyniok. Cambridge: Cambridge University Press, pp. 1–111. DOI: 10.1017/9781009025096.002 (cit. on p. 9).

Bilodeau, B. et al. (2024). "Impossibility theorems for feature attribution." In: *Proceedings of the National Academy of Sciences* 121.2, e2304406120 (cit. on p. 16).

Boden, M. A. (2016). *AI: Its nature and future*. Oxford: Oxford University Press (cit. on p. 5).

Bournez, O. and A. Pouly (2021). "A Survey on Analog Models of Computation." In: *Handbook of Computability and Complexity in Analysis*. Ed. by V. Brattka and P. Hertling. Cham: Springer International Publishing, pp. 173–226. DOI: 10.1007/978-3-030-59234-9_6 (cit. on p. 14).

Bradley, T.-D., J. Terilla, and Y. Vlassopoulos (2021). *An enriched category theory of language: from syntax to semantics*. arXiv: 2106.07890 [math.CT] (cit. on p. 13).

Bringsjord, S. and N. S. Govindarajulu (2024). "Artificial Intelligence." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta and U. Nodelman. Summer 2024. Metaphysics Research Lab, Stanford University (cit. on p. 5).

Caro, M. C. (12/2023). "From undecidability of non-triviality and finiteness to undecidability of learnability." In: *International Journal of Approximate Reasoning* 163, p. 109057. DOI: 10.1016/j.ijar.2023.109057. URL: https://arxiv.org/abs/2106.01382 (cit. on p. 10).

Colbrook, M. J., V. Antun, and A. C. Hansen (2022). "The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem." In: *Proceedings of the National Academy of Sciences* 119.12, e2107151119. DOI: 10.1073/pnas.2107151119. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2107151119. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2107151119 (cit. on p. 10).

Delétang, G. et al. (2023). *Neural Networks and the Chomsky Hierarchy*. arXiv: 2207.02098 [cs.LG] (cit. on p. 10).

Flasiński, M. (2016). *Introduction to Artificial Intelligence*. Cham: Springer. DOI: https://doi.org/10.1007/978-3-319-40022-8 (cit. on p. 7).

Grohs, P. and G. Kutyniok, eds. (2022). *Mathematical Aspects of Deep Learning*. Cambridge University Press. DOI: 10.1017/9781009025096 (cit. on p. 2).

Hensel, F., M. Moor, and B. Rieck (2021). "A Survey of Topological Machine Learning Methods." In: *Frontiers in Artificial Intelligence* 4. DOI: 10.3389/frai.2021.681108 (cit. on p. 12).

Hornik, K., M. Stinchcombe, and H. White (1989). "Multilayer feedforward networks are universal approximators." In: *Neural Networks* 2.5, pp. 359–366. DOI: https://doi.org/10.1016/0893-6080(89)90020-8 (cit. on p. 8).

Immerman, N. (2021). "Computability and Complexity." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University (cit. on p. 7).

Kratsios, A. (2021). "The Universal Approximation Property." In: *Annals of Mathematics and Artificial Intelligence* 89.435-469. DOI: https://doi.org/10.1007/s10472-020-09723-1 (cit. on p. 8).

Naitzat, G., A. Zhitnikov, and L.-H. Lim (2020). "Topology of Deep Neural Networks." In: *Journal of Machine Learning Research* 21.184, pp. 1–40. URL: http://jmlr.org/papers/v21/20-345.html (cit. on p. 12).

Roberts, D. A. and S. Yaida (2022). *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge: Cambridge University Press. DOI: 10.1017/9781009023405. URL: https://arxiv.org/abs/2106.10165 (cit. on p. 11).

Russell, S. J. and P. Norvig (2021). *Artificial Intelligence: A Modern Approach*. Pearson series in artificial intelligence. Harlow: Pearson (cit. on pp. 3, 5).

Shalev-Shwartz, S. and S. Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. A copy for personal use only at https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/copy.html. New York: Cambridge University Press (cit. on p. 8).

Shiebler, D., B. Gavranović, and P. Wilson (2021). *Category Theory in Machine Learning*. URL: https://arxiv.org/abs/2106.07032 (cit. on p. 13).

Šíma, J. and P. Orponen (2003). "General-Purpose Computation with Neural Networks: A Survey of Complexity Theoretic Results." In: *Neural Computation* 15, pp. 2727–2778 (cit. on p. 10).

Sterkenburg, T. F. and P. D. Grünwald (2021). "The no-free-lunch theorems of supervised learning." In: *Synthese* 199.3-4, pp. 9979–10015. DOI: 10.1007/s11229-021-03233-1 (cit. on p. 8).

Van Bekkum, M. et al. (2021). "Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases." In: *Applied Intelligence* 51, pp. 6528–6546. DOI: 10.1007/s10489-021-02394-3 (cit. on p. 13).