清華大學
Tsinghua University

Reading Materials for Machine Learning Theory

Angelica Aviles-Rivero
aviles-rivero@tsinghua.edu.cn

Version: December 24th, 2024

What to Read? Key Readings for Machine Learning Theory

Machine learning is the study of algorithms that allow machines to learn patterns and make predictions based on data. Formally, it explores how a system can generalise from observed samples to unseen instances, minimising error whilst balancing constraints such as computational complexity and data availability. In essence, **learning** is a mathematical process that seeks to identify a hypothesis function from a class of possible functions that performs well on unseen data.

The goal of machine learning is to **approximate an unknown function** that maps inputs to outputs, given a finite set of observations. For example, the task of predicting whether an email is spam can be described mathematically: we seek a function $h \in \mathcal{H}$, from a hypothesis class $\mathcal{H}$, such that $h(x) \approx y$ for input-output pairs (x, y) drawn from an unknown probability distribution D. The challenge is to identify a function that minimises error on new, unseen samples—not just the observed data. Central to the theory of learning is the concept of **generalisation**: how well a hypothesis learned from a training set applies to unseen instances. This requires us to formalise notions of risk, such as the expected error (or generalisation error), and develop algorithms that can efficiently search the hypothesis space to minimise it.

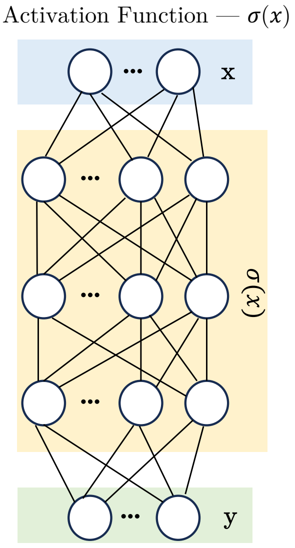
However, learning comes with fundamental limitations. **Overfitting**, for instance, occurs when a model performs well on the training data but poorly on unseen data. This leads us to trade-offs between model complexity and accuracy, captured mathematically through frameworks like Empirical Risk Minimisation (ERM), Structural Risk Minimisation (SRM), and regularisation techniques.

What to Expect from This Course

In this course, we will delve deeply into the mathematical underpinnings of machine learning. Our goal is not just to implement algorithms but to rigorously understand

why they work and when they are expected to fail. You will engage with mathematical concepts such as probability theory, optimisation, and linear algebra, as they are essential tools for analysing learning algorithms. Here is an of what you can expect:

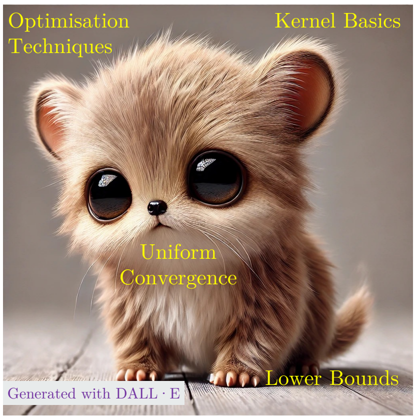
- **Introduction to Machine Learning Theory.** This part lays the groundwork by introducing what machine learning theory entails and why it is critical to modern science. We will discuss supervised and unsupervised learning paradigms, essential for understanding how models are trained and validated. Additionally, decision theory is introduced to frame predictions mathematically. This foundational knowledge helps clarify the scope and limitations of machine learning models and prepares students for deeper theoretical exploration.
- **Formal Foundations of Learning.** This part focuses on formalising the process of learning using the PAC (Probably Approximately Correct) framework. The PAC framework provides the mathematical tools to reason about whether a machine can learn a given task. VC-dimension is introduced to quantify the capacity of hypothesis classes and their ability to generalise. Understanding these topics is critical for assessing whether a learning algorithm is well-suited for specific tasks and data.


Activation Function — $\sigma(x)$

- **Empirical Risk Minimisation and Regularisation.** ERM is a fundamental principle that underpins many learning algorithms. This week explores how ERM minimises errors on training data and highlights its shortcomings, particularly overfitting. Regularisation techniques like ridge regression are introduced to mitigate overfitting by balancing model complexity and performance. The bias-variance trade-off is discussed as a central framework for navigating the trade-offs inherent in model design. These concepts are key to building reliable and effective machine learning systems.
- **Complexity Measures and Learnability.** We delve into advanced measures of complexity such as Rademacher complexities and uniform convergence. These tools help assess the ability of a model to generalise to unseen data. We explore Sauer's lemma and growth functions, which provide theoretical insights into how complexity affects learnability. These topics are essential for selecting models that balance expressiveness with generalisation.
- **Computational Complexity and Generalisation.** Understanding the computational limits of learning is crucial for developing efficient algorithms. During this part of the course, we introduce the No-Free-Lunch theorem, which illustrates the trade-offs

involved in making assumptions about learning tasks. PAC-Bayesian bounds are introduced to provide probabilistic guarantees for model generalisation. These insights are critical for understanding the practical feasibility of machine learning systems.

- **Advanced Generalisation Techniques.** This section expands on generalisation by introducing covering numbers, which refine our ability to bound generalisation errors. PAC-Bayesian analysis is revisited with a focus on applications to high-dimensional data, a critical challenge in modern machine learning. These techniques are vital for ensuring that models remain robust and reliable even in complex scenarios.
- **Optimisation Techniques.** Optimisation lies at the heart of training machine learning models. This part of the course introduces convex optimisation and stochastic gradient descent (SGD), focusing on their convergence properties in convex settings. Adaptive optimisation techniques, which are widely used in deep learning, are also discussed. Grasping these methods is crucial for implementing efficient and scalable machine learning algorithms.



Generated with DALL · E

- **Kernel Methods and High-Dimensional Data.** Kernel methods provide a powerful framework for learning non-linear patterns in data. This week introduces the representer theorem and discusses common kernels such as polynomial and Gaussian kernels. Practical applications in high-dimensional data, including support vector machines (SVMs), are explored. These methods are particularly relevant for tasks requiring robust non-linear decision boundaries.
- **Neural Networks.** Neural networks are highly relevant to modern machine learning. This part of our course focuses on single-layer networks, exploring their ability to approximate functions through the universal approximation theorem. Challenges such as overparameterisation are discussed alongside neural tangent kernels, which provide theoretical insights into the success of deep learning. Understanding these topics is key to leveraging neural networks effectively in practice.
- **Ensemble Learning and Online Learning.** Ensemble methods such as bagging and boosting are introduced as strategies to improve model accuracy and robustness. Online learning algorithms, including weighted majority approaches, are explored for their ability to adapt to sequential data. These methods are essential for building flexible and adaptive systems that can handle dynamic environments.
- **Probabilistic Methods.** Probabilistic reasoning is important for modelling uncertainty in machine learning. This week covers Bayesian inference and Gaussian processes, which allow for robust uncertainty estimation. PAC-Bayesian analysis is revisited, with

practical examples in model selection and evaluation. These topics are particularly relevant for applications requiring reliable confidence measures.

- **Structured Prediction.** Many tasks involve predicting structured outputs, such as sequences or rankings. We will explore structured SVMs, loss functions for structured data, and decoding techniques. Mastering these methods is critical for tackling real-world tasks that extend beyond simple classification or regression.
- **Overparameterisation and Implicit Bias.** Modern deep learning systems often rely on overparameterised models. We examine the double descent phenomenon and the implicit bias of gradient descent, which explain why these models can generalise effectively despite their complexity. Understanding these phenomena is crucial for designing effective deep learning architectures.
- **Advanced Optimisation Techniques.** Non-convex optimisation presents significant challenges in machine learning. In this section, we explore strategies for navigating non-convex landscapes and introduces adaptive methods for large-scale problems. Techniques for handling sparse and high-dimensional data are also discussed, with applications to modern learning problems.
- **Causal Inference and Reinforcement Learning.** The course concludes with causal inference, focusing on causal models and counterfactual reasoning for decision-making. Reinforcement learning theory is introduced to address sequential decision-making problems. These topics are essential for building systems that make informed decisions based on cause-effect relationships.

Reading Materials

This course is supported by several core texts that provide the essential theoretical and practical knowledge required for this course on machine learning theory. In addition to the main readings, extra materials will be recommended throughout the course for those who are curious to explore advanced topics and gain deeper insights. Below is a brief description of the primary materials we will use, along with supplementary texts available for further study.

Main Materials

★**Learning Theory from First Principles by** (Bach, 2024). This textbook offers a rigorous exploration of learning theory, emphasising derivation from first principles and linking theoretical insights to practical implementation. It covers essential topics in statistical learning, optimisation techniques, kernel methods, and overparameterised models, while discussing challenges such as adaptivity, trade-offs in approximation, estimation, and optimisation errors. Special topics include ensemble learning, online learning, and structured prediction, enriched by illustrative experiments and exercises. The text is tailored for theory-oriented students and researchers who seek to understand the mathematical foundations and algorithmic innovations underpinning modern machine learning.

What Content Is Expected to Be Learned? A general overview of the foundational concepts can be found in Chapter 1, providing students with a structured introduction to machine learning fundamentals. For those already familiar with the subject matter, you may choose to jump directly into the core theoretical chapters. **The main theoretical fundamentals expected to be acquired in-depth are covered in Chapters 2, 4, 5, 6, 7, and 14.** These chapters encompass topics such as supervised learning paradigms, empirical risk minimisation, statistical learning theory, kernel methods, and probabilistic generalisation frameworks, all critical for understanding learning theory. In addition, we will examine specific algorithms and advanced theoretical discussions to gain a broad understanding of their purpose and implications. **While a detailed study isn't required for these, general familiarity with their definitions and motivations is encouraged.** These are found in Chapters 9, 10, 12, 13, and 15, which include neural networks, ensemble learning, overparameterisation, structured prediction, and theoretical lower bounds.

📖 The authors of that (Bach, 2024) provide a copy for personal use, as indicated by the authors, at the following link.

★**Understanding Machine Learning: From Theory to Algorithms by** (Shalev-Shwartz and Ben-David, 2014). This book offers a well-rounded introduction to the theory and algorithms that form the foundation of modern machine learning. It begins by addressing fundamental questions about learning, including how learning can be formally defined and under what conditions it succeeds. The text delves into key concepts such as PAC learning, VC-dimension, and empirical risk minimisation, which underpin the theoretical side of the field. Alongside theory, the book introduces practical algorithms such as linear models, neural networks, and support vector machines, and explores optimisation methods like gradient descent and regularisation. Special attention is given to challenges such as overfitting, model selection, and evaluation strategies, equipping readers with the tools needed to build effective models.

What Content Is Expected to Be Learned? A general overview of the foundational concepts can be found in Chapter 1, offering students a structured introduction to machine learning fundamentals. For those already familiar with the subject matter, you may choose to jump directly into the core chapters. **The main theoretical fundamentals expected to be acquired in-depth are covered in Chapters 2, 3, 5, 6, 13, and 14**. These chapters provide the mathematical and theoretical grounding essential for this course, covering topics such as PAC learning, VC-dimension, risk minimisation, and key algorithmic frameworks. In addition, we will explore several particular algorithms to establish a general understanding of their definitions and motivations. **While a detailed study isn't required for these, a general knowledge of their purpose and application is important.** These are found in Chapters 9, 15, 18-20, and 22-23.

📖 The authors of that (Shalev-Shwartz and Ben-David, 2014) provide a copy for personal use, as indicated by the authors, at the following link.

★Convex Optimisation: Algorithms and Complexity by (Bubeck et al., 2015). This material offers a comprehensive introduction to convex optimisation with a particular focus on algorithms and their complexities. It begins by addressing fundamental aspects of convexity, such as the properties of convex functions and sets, and explains why convexity plays a central role in optimisation. The text also seeks to cover essential algorithms, including gradient descent, cutting plane methods, and stochastic optimisation, highlighting their convergence rates and computational feasibility. With a strong emphasis on both the theoretical underpinnings and practical implementation, this text is particularly valuable for optimisation and machine learning. The structured presentation makes it a great resource for the course, providing deeper insights into optimisation techniques critical for machine learning models.

What Content Is Expected to Be Learned? To gain a solid general understanding from this reading, students should focus on **Chapter 1 (a concise and very short overview) and Section 3.2**. These sections cover the foundational concepts and core methods sufficiently for a comprehensive introduction without delving into overly detailed analysis. This targeted reading is designed to equip students with essential insights into convex optimisation's role in machine learning and the main algorithms relevant to this course.

📖 The pre-print version of this material is available online in this link.

★Practical Implementation Resources. Whilst this course focuses on machine learning theory, there will also be a practical component to reinforce your understanding. For this, homework exercises will primarily involve **Python-based implementations**. We encourage you to explore and become familiar with Python's scikit-learn library, as it will be the main tool used for practical assignments.

🌐 For reference and guidance, please use the Scikit-learn project's official documentation in here.

Supplementary Materials

The following materials are supplementary and intended for students who are curious to gain deeper insights into the theory and practical aspects of machine learning. These resources are not part of the core curriculum but are provided to support those interested in exploring the subject further.

★Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation by (Belkin, 2021). It explores foundational mathematical concepts related to deep learning, with particular focus on interpolation and over-parameterisation. The work is an attempt to bridge the gap between the theoretical underpinnings and practical success of deep learning models, which have outpaced traditional learning theory.
📖 The pre-print version of this material is available online in this link.

★High-dimensional probability: An introduction with applications in data science by Vershynin (2018). This supplementary material reading provides a rigorous exploration of probability theory specifically tailored for high-dimensional contexts, which are essential for understanding the theoretical foundations of machine learning. The book introduces essential concepts like concentration inequalities, random matrices, high-dimensional distributions, and random projections—all of which are central to the mathematical understanding of machine learning algorithms, particularly in high-dimensional and over-parameterised models. These topics align well with understanding how machine learning models generalise, optimise, and handle high-dimensional data, making it a valuable resource for students in this machine learning theory course.
📖 You can access a free draft of this material here, *provided it is used only for personal and classroom needs as indicated by the author.*

Bibliography

Francis Bach. *Learning theory from first principles.* MIT press, 2024. 6, 7

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021. 9

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 2015. 8

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014. 7, 8

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. 9