Note: This syllabus may change over the course of the semester.

For the individual topics, we may sometimes provide links to Related Material:

- [Ma] Stanford CS229M https://github.com/tengyuma/cs229m_notes/blob/main/master.pdf
- [Bach] Learning Theory from First Principles (draft: https://www.di.ens.fr/~fbach/ltfp_book.pdf)
- [SSBD] Understanding Machine Learning: From Theory to Algorithms https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/
- [MRT] Foundations of Machine Learning https://cs.nyu.edu/~mohri/mlbook/

Note: We provide these references so that you can check out extra background if you want. Reading them is not part of the required workload for the class.

As applications of the techniques we're developing, we're planning to especially cover the following research papers:

- Edelman et al, ICML 2022: https://arxiv.org/abs/2110.10090 (preprint), https://proceedings.mlr.press/v162/edelman22a/edelman22a.pdf (ICML version)
- Lotfi et al, ICML 2024: https://arxiv.org/abs/2312.17173
- Huang et al, ICLR 2025: https://arxiv.org/abs/2410.02140

Schedule

Date (DD.MM.2026)		**Topic**
09.04.2026 Michael	Thu	Intro, Preliminaries Basic Concepts: Empirical Risk Minimization Probability Background Material: ● SSBD: Section 2
13.04.2026 Michael	Mo	Markov's Inequality Chebyshev's Inequality A first quantitative generalization bound Material: ● Ma: Section 3.2 ● Bach: Exercises 1.9, 1.10

16.04.2026 Michael	Thu	Hoeffding's Inequality Refined bound for finite classes Material: ● Please look at Homework 1 if you haven't done so ● https://en.wikipedia.org/wiki/Hoeffding%27s_inequality ● Ma: Section 3.3, 4.1.1, 4.2 ● Bach: Section 1.2.1, 4.2, 4.4.3 ● SSBD: Section 4.2
20.04.2026 Yash	Mo	(1) Probability Essentials (Random Variables, Expectation, Independence, Conditional Expectation) ● Jensen's Inequality (2) Vector & Matrix Norms ● Cauchy Schwarz Inequality
23.04.2026 Michael	Thu	(1) Refined bound for finite classes (continued) Material: ● Ma: Theorem 4.1 ● Bach: Section 4.4.3 "Finite Number of Models" (2) *(We added this to give an explicit example for our results so far, and postponed Rademacher complexity to Monday)* Example: Learning of Finite-State Automata Material: ● https://en.wikipedia.org/wiki/Deterministic_finite_automaton ● *If you want to learn more, you can check out MRT Chapter 16. This is not required for the class.* ● *This is the recent paper on extracting automata from transformers that was mentioned in class: https://openreview.net/forum?id=L8SMNWsxfK* (3) Homework 2 is released (around Friday).
27.04.2026 Michael	Mo	(1) Discussing results of feedback survey. (2) Discussing solution for Homework 1 (3) Rademacher Complexity: Definition and Symmetrization. Material:

		• Ma: Section 4.4, especially 4.4.1-4.4.3, and Theorem 4.13 • Bach: Eq. 4.12, Proposition 4.2 • SSBD: Eq 26.4, 26.5, Lemma 26.2, Lemma 26.3
30.04.2026 Michael	Thu	(1) Lipschitz Composition. In-Expectation Generalization Bound for Linear Classes. Material: • Ma: Theorem 5.5 • Bach: Sections 4.5.2, 4.5.3, 4.5.4 • SSBD: Lemma 26.9, Lemma 26.10 <u>Notes:</u> 1. For the behavior of Rademacher complexity under Lipschitz functions, which we stated in class but didn't prove, you can read the proofs in: • Bach Prop. 4.3 • SSBD Lemma 26.9 if you're interested. In the end, however, the Lipschitz composition of *covering numbers* will be more important for us; we'll prove it soon after. 2. Our in-expectation bounds on the excess risk can be converted into high-probability bounds. You can find this in: • Ma: Theorem 4.18 • SSBD: Theorem 26.5 if you're interested.
04.05.2026 Michael	Mo	(1) Massart's Finite Class Lemma *Note: In [Ma], this is stated as Proposition 4.6.1 without proof as part of the section on Covering Numbers. As promised in the lecture, you can find the proof under "Materials" in CMS.* Material: • SSBD Lemma 26.8 (Massart Lemma) (2) Covering Numbers Material: • Ma: Section 4.6, beginning of “Covering number upper bounds Rademacher complexity”, up to and including Theorem 4.25. • Bach: Section 4.4.4 • SSBD: Section 27.1, especially Definition 27.1

		Note: We will treat (1) and (2), and perhaps (3). We'll postpone Dudley, which previously was listed here, to later. (3) Lipschitz composition for covering numbers: ● Ma: Lemma 4.29 ● Bach: Proposition 4.3 ● SSBD: Lemma 27.3 *We will release mock exam questions during this week.*
07.05.2026 Yash	Thu	(1) Transformers 101 (setting up notation for theoretical papers) Material: *Note that there are many different notation conventions for the components of transformers.* ● Section 3 in Edelman et al, ICML 2022: https://proceedings.mlr.press/v162/edelman22a/edelman22a.pdf (2) Homework 3 (about Transformers and covering numbers) is released (around Friday). If you missed the class, going through Section 3 of the Edelman paper should suffice. If you are curious about Transformers and just want to look at examples of how theoretical papers treat Transformers, you can go through the following excellent resource that teaches how to construct Transformers to solve toy problems by hand. https://openreview.net/pdf?id=sPshCSvDrX
11.05.2026 *asynchronous class*	Mo	*asynchronous class: practice questions*
14.05.2026	Thu	Gesetzlicher Feiertag: Christi Himmelfahrt No lecture!
18.05.2026	Mo	(1) Solution for Homework 2 (2) Lipschitzness of Transformer Heads Material: ● Please revisit the material about vector norms from April 20.

		• Edelman et al, ICML 2022: https://proceedings.mlr.press/v162/edelman22a/edelman22a.pdf ○ Lemma 4.3 (Lipschitzness of attention head) • If you're curious, Lipschitzness properties of attention have been shown in many papers in various ways and with different statements. This paper may be interesting from a real-world perspective: https://arxiv.org/abs/2505.10606 . Some other references include: https://arxiv.org/abs/1906.06755 ; https://arxiv.org/abs/2402.09963 .
21.05.2026 Yash	Thu	(1) FFN Constructions : Cancelling out Residual Streams, Identity without Residuals, Min / Max. (2) Statement of the Universal Approximation Theorem – Cybenko'89 (3) Attention Variants – Softmax Attention vs Hard Attention. (4) Homework 4 (Transformers & sparse functions) is released (around Friday; due: June 10). Material (for Transformer Constructions): https://openreview.net/pdf?id=sPshCSvDrX Sections 4 (Intro), 4.2, 4.3, 4.4, Section 5 (Intro)
25.05.2026	Mo	Gesetzlicher Feiertag: PfingstMo No lecture!

28.05.2026
Michael
	Thu	(1) Solution for Homework 3
		
		(2) Covering Number-based Generalization for Transformers
		
		Material:
		• Edelman et al, ICML 2022: https://proceedings.mlr.press/v162/edelman22a/edelman22a.pdf
		  ○ Lemma 4.3 (Lipschitzness of attention head)
		  ○ passing to Theorem 4.2 (covering-based statement)
		  ○ passing to Corollary 4.5 (covering number of attention heads)
		  ○ passing to Theorem 4.7 (covering number of transformer block)

		(3) Dudley entropy integral *We will probably only cover the statement but not the proof, which you can find in these materials if you want. It's an extension of the proof from May 4.* Material: ● Ma: Section 4.6, subsection “Chaining and Dudley’s theorem”; Lemma 4.28 ● SSBD: Section 27.2, especially Lemma 27.4 and Lemma 27.5 (4) Covering number bound for linear functions *We will only cite the result from this paper, as used by Edelman et al. You can read the material if you want to learn more.* Material (*just if you're curious*): ● Zhang 2022: https://www.jmlr.org/papers/volume2/zhang02b/zhang02b.pdf (4) Generalization Bound for Transformers (putting together all results we have obtained so far)
01.06.2026 Yash	Mo	(1) Moving from In-Distribution Generalization to Length Generalization (2) Huang et al’2025 Paper Motivation, goal and the impact of such theory on LLMs. (3) Huang et al’2025 Informal Run through of the whole paper. Introducing Limit Transformers, C-RASP Material: Huang et al, ICLR 2025: https://arxiv.org/abs/2410.02140 (Main paper – first 10 pages, detailed formalization & proofs will be discussed in subsequent classes).
04.06.2026	Thu	Gesetzlicher Feiertag: Fronleichnam No lecture!
08.06.2026 Yash	Mo	(1) C-RASP & C-RASP[Periodic, Local] Formal Definition (2) Example Programs ● Majority ● Dyck-1

		• Unique Right Retrieval (~Induction Head) (3) Introducing Transformer Parameterization (4) Intuition behind the Product Functions Parameterization. Material. 1. Huang et al, ICLR 2025: https://arxiv.org/abs/2410.02140 • Notation: Section 2 • Product Function Intuition: Section 3.1 • C-RASP Definition: Section 4.2 • C-RASP Example Programs: Theorem 10 • Intuition behind some programs being outside CRASP: Lemma 11. 2. Jobanaputra et al. NeurIPS 2025: https://proceedings.neurips.cc/paper_files/paper/2025/file/894403f9604374a7a003063e480f65b9-Paper-Conference.pdf • Generation with C-RASP & Unique Right Induction: Lemma 4 (Appendix A.2)
11.06.2026 Yash	Thu	(1) Hypothesis Class, Regularizer, Inference Procedure (2) Limit Transformers (3) Main Length Generalization Theorem Homework 5 is released (around Friday). Material. 1. Huang et al, ICLR 2025: https://arxiv.org/abs/2410.02140 • Theoretical Framework: Section 3
15.06.2026 Yash	Mo	(1) Communication Complexity Argument (Just statement, no proof). (2) Proof of the Main Length Generalization Theorem started. Material. • Huang et al, ICLR 2025: https://arxiv.org/abs/2410.02140 • Definition 50, Theorem 12, Corollary 13. • Theorem 7 proof high level idea. • Lemma 17 proof started (to be continued in a later class.)

18.06.2026 Yash	Thu	(1) Proof of Lemma 17 from Huang et al continued.
22.06.2026 Yash	Mo	(1) Solution for Homework 4 (2) Solutions to the Practice questions shared on the 4th May' 2026. Questions 1 to 15.
25.06.2026 Yash	Thu	(2) Solution for Homework 5. (2) Solutions to the Practice questions shared on the 4th May' 2026. Questions 16 to 35.
29.06.2026 Michael	Mo	(1) Recap of the big picture: Covering Numbers vs Length Generalization as approaches to Transformer learning theory. Challenges and open questions for the field. *Optional readings if you're interested*: ● Challenges of Uniform Convergence ○ Zhang et al, ICLR 2017: https://arxiv.org/abs/1611.03530 ○ Wilson, ICML 2025: https://arxiv.org/abs/2503.02113 ● Open Questions for Length Generalization ○ Izzo et al, ICLR 2026: https://arxiv.org/abs/2510.27015 ○ Yang et al, ICML 2026: https://arxiv.org/abs/2603.02238 (2) Compression-Based Bounds / PAC Bayes *Not directly relevant to the exam!* Material: ● Lotfi et al, ICML 2024: https://arxiv.org/abs/2312.17173 ● We'll mainly be concerned with Theorem A.1 (prior-weighted Hoeffding bound)
02.07.2026 Michael	Thu	Practise Questions for the Exam/ Revision/ Doubts?
06.07.2026 *Asynchronous class*	Mo	*Mock Exam*

09.07.2026 Michael	Thu	Mock Exam solutions
13.07.2026 Yash	Mo	EXAM Normal time and place of the lecture.
28.09.2026	Mo	RE-EXAM 14:00 to 16:00 Location: Building C7.2, Room -1.05 (minus 1.05)