IEMS 402: Statistical Learning

Northwestern University - Winter 2025 - Syllabus

This course provides foundational and advanced concepts in statistical learning theory, essential for analyzing complex data and making informed predictions. Students will delve into both asymptotic and non-asymptotic analyses of machine learning algorithms, addressing critical challenges such as model bias, variance, and robustness in uncertain environments. Toward the end of the course, students will apply these principles to modern machine learning contexts, including the scaling laws/benign overfitting of deep learning, generative AI, and language models. (e.g. Neural Tagent Kernel, Mean-Field Limit of Neural Network and In-context Learning)

Course Homepage: https://2prime.github.io/teaching/2025-Statistical-Learning
Professor: Prof. Yiping Lu **Lecture:** MW 2:00-3:20
Email: yiping.lu@northwestern.edu **Office Hour:** W 3:30-4:30 M237
If you wish to contact me via email, kindly include the tag "[IMES 402]" in the subject line. This will help ensure that I do not overlook your message. Better way to approach me is using campuswire. ***Please always utilize Campuswire instead of emailing*** – this helps centralize conversations between us

Grading Distribution: Problem Sets (15%) + Exams (80%) + Scribe Note (5%)

Problem Sets (15%)
Problem sets will be posted to Course homepage and Gradescope. A final draft of your writeup *must be submitted in PDF*. It should include your thought process, calculations showing all work and citations of theorems and definitions used. For all problem sets, you are highly encouraged to work in groups. Problem sets will be due by midnight on Fridays with only one late assignment (up to 12 hours) accepted per student. Moreover, *your grade is computed by max(HW1,HW8)+max(HW2,HW3)+max(HW4,HW5)+max(HW6,HW7).*
No further extensions or considerations will be given.
[Homework 1] [Homework 2] [Homework 3] [Homework 4] [Homework 5] [Homework 6] [Homework 7] [Homework 8]

For your first late assignment within 12 hours after the deadline (as indicated on Gradescope), no point deductions. All subsequent assignments submitted within 12 hours after the deadline will convert to a zero at the end of semester. In all cases, work submitted 12 hours or more after the deadline will not be accepted.

Exams (70%)
There are 2 exams in the course as indicated on the schedule – a midterm and a noncumulative final. Please do not inquire about curves, grades will be assigned equitably. More information about the weight distribution will be communicated throughout the semester. The exam will strictly follow the concepts/techniques appearing in the [Example Midterm] [Example Final].

Scribe Notes (5%)
You will be responsible for the scribe notes of one lecture. Students will sign up (possibly in pairs) to scribe each lecture; students have 72 hours to type up the scribe notes. You will be provided a draft version of the lecture note. The lecture notes will be available to edit by all on Overleaf; you're encouraged to add to and improve the notes for any week, regardless of whether it is the week you are responsible for. Overleaf allows us to see who made which edits; your contributions across all 18 lectures of the class will be counted. Full credit will only be given to students who produced complete, useful notes!
Some comments on this: In a lecture, the instructor usually only writes a subset of what they are communicating on the board. Good scribe notes are not just a latex transcript of what the instructor wrote; they fill in the gaps. For this reason, producing scribe notes requires a thorough understanding of the material, which often entails some additional reading of the supplementary texts.

Book:

- Learning Theory from First Principles Francis Bach [LTFP] https://www.di.ens.fr/~fbach/ltfp_book.pdf
- Asymptotic Statistics: https://www.cambridge.org/core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D [AS]

Each lecture will be scribed by a student, and the scribe notes will be available for all to edit on Overleaf. You will be added to the Overleaf project in the first week of the quarter.

Relevant Course:

- Stanford Stats 300b: https://web.stanford.edu/class/stats300b/
- Stanford CS229T: https://web.stanford.edu/class/stats214/
- Stanford Stats 311: https://web.stanford.edu/class/stats311/lecture-notes.pdf
- Berkeley Stats 241: https://www.stat.berkeley.edu/~bartlett/courses/2014spring-cs281bstat241b/
- Berkeley Stats 241B: https://github.com/ryantibs/statlearn-s24/tree/main
- CMU Stat705 : https://www.stat.cmu.edu/~larry/=stat705/
- CMU 10-072: https://www.stat.cmu.edu/~ryantibs/statml/
- UW Madison CS 839 : https://pages.cs.wisc.edu/~yudongchen/cs839_sp22/
- UIUC ECE598YW: http://www.stat.yale.edu/~yw562/teaching/598/index.html
- MIT IDS.160/9.521/18.656/6.S988 https://www.mit.edu/~rakhlin/courses/mathstat/rakhlin_mathstat_sp22.pdf
- Umich EECS598 https://web.eecs.umich.edu/~cscott/past_courses/eecs598w14/

Preliminary: Review Note

- Calculus, Linear Algebra
- Probability and Statistics (IEMS 302): Strong Law of Large Numbers, Central Limit Theorem, Big-O, little-o notation
- Optimization (Convex duality) (http://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf)

Advanced Textbook:

- Algorithmic Aspects of Machine Learning Ankur Moitra
- https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.html

- https://www.cambridge.org/core/books/highdimensional-statistics/8A91ECEEC38F46
- DAB53E9FF8757C7A4E

- Growth mindset holds that if you work at something, you can improve.
- Mistakes are an opportunity to learn. There is no failure. The only failure is not improving.

About your instructor

I (Dr Yiping Lu) am Assistant Professor of Industrial Engineering & Management Sciences joining Northwestern University in 2024. Before that, I worked as a Courant Instructor at NYU for one year. From 2019 to 2023, I was a Ph.D. student at Stanford University where I obtained my degree in Applied Mathematics, emphasis on Machine Learning (AI) and Numerical analysis (using computers to solve equations). My research is about using AI to solve hard physics, industrial engineering and system management problems. Our department (IEMS) at Northwestern University is working on how to make important decisions using data based on linear algebra and statistics. My native language is Mandarin, and I also speak Japanese.

Accommodations

I am happy to provide accommodations, understanding that they may be necessary for student success. Students who may need academic accommodation based on the impact of a disability must initiate the request with the AccessibleNU. Students should contact AccessibleNU as soon as possible since timely notice is needed to coordinate accommodations.

Chapter 0: What is Machine/Statistical Learning

0.1 Course Overview

[LTFP] Section 2

- Supervised Learning, Designing loss function via max log likelihood, Surrogate Loss
- Overfitting, Bias-variance trade-off, Scaling Law
- Machine Learning Error = Generalization Error + Approximation Error + Optimization Error
- Challenges of Machine Learning Theory:
    - Curse of dimensionality
    - Over-Parameterization/Interpolation
    - Non-Convex Optimization (landscape, sharpness)
    - Distribution Mismatch

Suggested Reading:

- Zhang C, Bengio S, Hardt M, et al. Understanding deep learning (still) requires rethinking generalization. Communications of the ACM, 2021, 64(3): 107-115.
- Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.
- Nakkiran P, Kaplun G, Bansal Y, et al. Deep double descent: Where bigger models and more data hurt. Journal of Statistical Mechanics: Theory and Experiment, 2021, 2021(12): 124003.
- Curth A, Jeffares A, van der Schaar M. A U-turn on double descent: Rethinking parameter counting in statistical learning. Advances in Neural Information Processing Systems, 2024, 36.
- Rethinking Conventional Wisdom in Machine Learning: From Generalization to Scaling L Xiao arXiv preprint arXiv:2409.15156
- Mallinar N, Simon J, Abedsoltan A, et al. Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. Advances in Neural Information Processing Systems, 2022, 35: 1182-1195.
- Li H, Xu Z, Taylor G, et al. Visualizing the loss landscape of neural nets. Advances in neural information processing systems, 2018, 31.
- Keskar N S, Mudigere D, Nocedal J, et al. On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836, 2016.
- Degree of Freedom: https://www.stat.cmu.edu/~ryantibs/advmethods/notes/df.pdf

Advanced Reading:

- Hastie T, Montanari A, Rosset S, et al. Surprises in high-dimensional ridgeless least squares interpolation. Annals of statistics, 2022, 50(2): 949.
- Bartlett P L, Long P M, Lugosi G, et al. Benign overfitting in linear regression. Proceedings of the National Academy of Sciences, 2020, 117(48): 30063-30070.
- Ge R, Lee J D, Ma T. Matrix completion has no spurious local minimum. Advances in neural information processing systems, 2016, 29.

0.2 Bias-Variance Trade-off Examples using Kernel Smoothing

[LTFP] Section 6 In this lecture, we aim to understand the Bias and Variance Trade-off using kernel smoothing as an example

Suggested Reading:

- Nonparametric regression: https://www.stat.cmu.edu/~ryantibs/statml/lectures/nonpar.pdf

Advanced Reading:

- Xing Y, Song Q, Cheng G. Benefit of interpolation in nearest neighbor algorithms. SIAM Journal on Mathematics of Data Science, 2022, 4(2): 935-956.
- Chhor J, Sigalla S, Tsybakov A B. Benign overfitting and adaptive nonparametric regression. Probability Theory and Related Fields, 2024: 1-32.

0.3 (Not Required) Paradigm of Learning: Unsupervised Learning, Semi-supervised learning and Generative AI

- Spectral Clustering, t-sne, Infomax and self-supervised learning
- Connection between the three methods

Suggested Reading:

- Zhou X, Belkin M. Semi-supervised learning by higher order regularization Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011: 892-900.
- Linderman G C, Steinerberger S. Clustering with t-SNE, provably. SIAM journal on mathematics of data science, 2019, 1(2): 313-332.
- Hjelm R D, Fedorov A, Lavoie-Marchildon S, et al. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670, 2018.
- HaoChen J Z, Wei C, Gaidon A, et al. Provable guarantees for self-supervised deep learning with spectral contrastive loss. Advances in Neural Information Processing Systems, 2021, 34: 5000-5011.

Advanced Reading:

- X. Cheng and N. Wu. "Eigen-convergence of Gaussian kernelized graph Laplacian by manifold heat interpolation". Applied and Computational Harmonic Analysis, 61, 132-190 (2022)
- Cai T T, Ma R. Theoretical foundations of t-sne for visualizing high-dimensional clustered data[J]. Journal of Machine Learning Research, 2022, 23(301): 1-54.

Chapter 1: Basic of Machine Learning 1: Asymptotic Theory

1.1 Asymptotic normality,

- Delta method, Higher Order Delta Method, implicit delta method, asymptotic normality [AS Chap 3,4]
- Inverse function theorem, Moment method,

1.2 Influence function

- M-Estimator, Fisher Information [AS Chap 5]
- Influence function
- Cramer-Rao Bound and Con of Unbiased Estimators (James-Stein estimator)

Required Reading:

- Koh P W, Liang P. Understanding black-box predictions via influence functions International conference on machine learning. PMLR, 2017: 1885-1894. (ICML 2017 best paper)
- Basu S, Pope P, Feizi S. Influence functions in deep learning are fragile.

Chapter 2: Basic of Machine Learning 2: Non-asymptotic Theory

2.1 Hoeffding, Chernoff Bounds

Hoeffding Inequality, Chernoff Bound
Application: Johnson-Lindenstrauss and high-dimensional embedding

2.2 Uniform Laws and Empirical Process Theory

Uniform Laws, Symmetrization, Rademacher Complexity, [LTFP] Section 4.5
Advanced Reading:

- Kur, Gil, et al. Minimum Norm Interpolation Meets The Local Theory of Banach Spaces. Forty-first International Conference on Machine Learning.
- Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. Neurips 2018.
- Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. The Annals of Statistics, 46(6A):2747–2774, 2018

2.3 Covering Numbers and Dudley's Theorem

Ledoux-Talagrand Contraction Principle, Dudley’s theorem, localized complexity, non-parametric least square
Advacned Reading:

- Kur G, Putterman E, Rakhlin A. On the variance, admissibility, and stability of empirical risk minimization. Advances in Neural Information Processing Systems, 2024, 36.
- Xu, Yunbei, and Assaf Zeevi. "Towards optimal problem dependent generalization error bounds in statistical learning theory." Mathematics of Operations Research (2024).

2.4 (Not Required) Generalization Theory of Neural Network

Other Ideas of Generalization: PAC-Bayes, Algorithm Stability, …
Suggested Reading:

- Bartlett, P.L., 1998. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. IEEE transactions on Information Theory, 44(2), pp.525-536
- Bartlett, P.L., Foster, D.J. and Telgarsky, M., 2017, December. Spectrally-normalized margin bounds for neural networks. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6241-6250)
- Jiang Y, Neyshabur B, Mobahi H, et al. Fantastic generalization measures and where to find them. arXiv preprint arXiv:1912.02178, 2019.
- Jiang Y, Nagarajan V, Baek C, et al. Assessing generalization of SGD via disagreement. arXiv preprint arXiv:2106.13799, 2021.

Chapter 3: Application of Machine Learning

3.1 Robust Learning, Distributionally Robust Learning and Distribution Shifts

Suggested Reading:

- https://hsnamkoong.github.io/assets/pdf/LiuWaCuNa24-slides.pdf
- Duchi J, Namkoong H. Variance-based regularization with convex objectives. Journal of Machine Learning Research, 2019, 20(68): 1-55. (Neurips 2017 Best paper)
- Geirhos R, Jacobsen J H, Michaelis C, et al. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2020, 2(11): 665-673.
- Sagawa S, Koh P W, Hashimoto T B, et al. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.

Further Reading:

- Stanford CS329D Machine Learning Under Distribution Shift: https://thashim.github.io/cs329D/
- Arjovsky M, Bottou L, Gulrajani I, et al. Invariant risk minimization. arXiv:1907.02893, 2019.

3.2 Reproducing Kernel Hilbert Space and Neural Tangent Kernel

3.2.1 Reproducing Kernel Hilbert Space and Gaussian Process

[LTFP] Section 7

Advacned Reading:

- Haas M, Holzmüller D, Luxburg U, et al. Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension[J]. Advances in Neural Information Processing Systems, 2024, 36.
- Schölpple M, Steinwart I. Which Spaces can be Embedded in Reproducing Kernel Hilbert Spaces? arXiv preprint arXiv:2312.14711, 2023.
- Lu Y, Lin D, Du Q. Which Spaces can be Embedded in Lp-type Reproducing Kernel Banach Space? A Characterization via Metric Entropy. arXiv preprint arXiv:2410.11116, 2024.

3.2.2 (Not Required) Mean-field and Kernel Approximation of Neural Network

Suggested Reading:

- [LTFP] Section 12.3, 12.5
- Bach F. Breaking the curse of dimensionality with convex neural networks. Journal of Machine Learning Research, 2017, 18(19): 1-53.
- Chizat L, Oyallon E, Bach F. On lazy training in differentiable programming. Advances in neural information processing systems, 2019, 32.

Advanced Reading:

- Chizat L, Bach F. On the global convergence of gradient descent for over-parameterized models using optimal transport. Advances in neural information processing systems, 2018, 31.

- Yang G, Hu E J, Babuschkin I, et al. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. arXiv preprint arXiv:2203.03466, 2022.
    - Hayou S, Ghosh N, Yu B. Lora+: Efficient low rank adaptation of large models. arXiv preprint arXiv:2402.12354, 2024.
    - Ishikawa S, Karakida R. On the Parameterization of Second-Order Optimization Effective Towards the Infinite Width. arXiv preprint arXiv:2312.12226, 2023.

Chapter 4: Advanced Topics in Machine Learning

4.1 (Not Required) Implicit Bias

Suggested Reading:

- [LTFP] Section 12.1

Advanced Reading:

- https://www.stat.berkeley.edu/~ryantibs/statlearn-s24/lectures/ridgeless.pdf

4.2 Topics of Large Language Model: In-context Learning, Chain-of-Thoughts, Alignment

In-context Learning, Chain-of-thoughts and Circuit Theory, Alignment of AI

Advanced Reading:

- Kim J, Nakamaki T, Suzuki T. Transformers are minimax optimal nonparametric in-context learner. arXiv preprint arXiv:2408.12186, 2024.
- Von Oswald J, Niklasson E, Randazzo E, et al. Transformers learn in-context by gradient descent International Conference on Machine Learning. PMLR, 2023: 35151-35174.
- Shen L, Mishra A, Khashabi D. Do pretrained Transformers Really Learn In-context by Gradient Descent?. arXiv preprint arXiv:2310.08540, 2023.
- Giannou A, Yang L, Wang T, et al. How Well Can Transformers Emulate In-context Newton's Method?. arXiv preprint arXiv:2403.03183, 2024.
- Feng G, Zhang B, Gu Y, et al. Towards revealing the mystery behind chain of thought: a theoretical perspective. Advances in Neural Information Processing Systems, 2024, 36.
- Lee J, Xie A, Pacchiano A, et al. Supervised pretraining can learn in-context reinforcement learning. Advances in Neural Information Processing Systems, 2024, 36.

4.3 Optimal Transport and Sampling

4.1.1 Basics of Optimal Transport

Optimal Transport and It's Dual, Statistics and Computation of Optimal Transport distance

Suggested Reading:

- Si N, Blanchet J, Ghosh S, et al. Quantifying the empirical Wasserstein distance to a set of measures: Beating the curse of dimensionality. Advances in Neural Information Processing Systems, 2020, 33: 21260-21270.
- Computational Optimal Transport: https://arxiv.org/abs/1803.00567

Advanced Reading:

- Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems, 2013, 26.
- Goldfeld Z, Greenewald K. Gaussian-smoothed optimal transport: Metric structure and statistical efficiency International Conference on Artificial Intelligence and Statistics. PMLR, 2020: 3327-3337.
- Li J, Tang J, Kong L, et al. A convergent single-loop algorithm for relaxation of gromov-wasserstein in graph data. arXiv preprint arXiv:2303.

4.1.2 (Not Required) Sampling using Optimal Transport Gradient Flow, Diffusion Model

Sampling as optimization in Optimal Transport geometry, Diffusion model, Generative AI

Suggested Reading:

- Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. SIAM journal on mathematical analysis, 29(1):1–17, 1998.
- Liu Q, Wang D. Stein variational gradient descent: A general purpose bayesian inference algorithm. Advances in neural information processing systems, 2016, 29.
- Chen A Y, Sridharan K. From Optimization to Sampling via Lyapunov Potentials. arXiv preprint arXiv:2410.02979, 2024.
- Song Y, Sohl-Dickstein J, Kingma D P, et al. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, 2020.
- Lipman Y, Chen R T Q, Ben-Hamu H, et al. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022.
- Albergo M S, Boffi N M, Vanden-Eijnden E. Stochastic interpolants: A unifying framework for flows and diffusions. arXiv preprint arXiv:2303.08797, 2023.

Advanced Reading:

- Log-concave sampling: https://chewisinho.github.io/main.pdf
- (extremely hard and fun) Query lower bounds for log-concave sampling. Sinho Chewi, Jaume de Dios Pont, Jerry Li, Chen Lu, and Shyam Narayanan. August 2024.

Make-up Paper Reading:

Hardness of Learning:

- Xu, Yunbei, and Assaf Zeevi. "Towards optimal problem dependent generalization error bounds in statistical learning theory." https://arxiv.org/abs/2011.06186
- Cutler, Joshua, Mateo Díaz, and Dmitriy Drusvyatskiy. "The radius of statistical efficiency." arXiv:2405.09676 (2024). https://arxiv.org/abs/2405.09676
- Asymptotic normality and optimality in nonsmooth stochastic approximation Damek Davis, Dmitriy Drusvyatskiy, Liwei Jiang https://arxiv.org/abs/2301.06632
- Damek Davis and Dmitriy Drusvyatskiy. Graphical convergence of subgradients in nonconvex optimization and learning. arXiv:1810.07590, 2018.
- Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. Neurips 2018.
- Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. The Annals of Statistics, 46(6A):2747–2774, 2018.
- Kur, Gil, et al. Minimum Norm Interpolation Meets The Local Theory of Banach Spaces. Forty-first International Conference on Machine Learning.

Sampling and Diffusion Model:

- Karimi M R, Hsieh Y P, Krause A. Sinkhorn Flow as Mirror Flow: A Continuous-Time Framework for Generalizing the Sinkhorn Algorithm https://arxiv.org/abs/2311.16706
- Ye He, Alireza Mousavi-Hosseini, Krishnakumar Balasubramanian, and Murat A. Erdogdu (2024). A separation in Heavy-tailed Sampling: Gaussian vs. Stable Oracles for Proximal Samplers. NeurIPS 2024. https://arxiv.org/abs/2405.16736
- Albergo M S, Vanden-Eijnden E. NETS: A Non-Equilibrium Transport Sampler. arXiv:2410.02711, 2024.https://arxiv.org/abs/2410.02711
- Montanari A. Sampling, diffusions, and stochastic localization. arXiv:2305.10690Montanari A, Wu Y. Posterior sampling from the spiked models via diffusion processes. arXiv preprint arXiv:2304.11449, 2023.
- Chen Y, Eldan R. Localization schemes: A framework for proving mixing bounds for Markov chains FOCS 2022.
- Bruna J, Han J. Posterior sampling with denoising oracles via tilted transport. arXiv preprint arXiv:2407.00745, 2024.

Deep Learning Theory

- Understanding Optimization in Deep Learning with Central Flows https://arxiv.org/pdf/2410.24206
- Weak-to-Strong Generalization and Scaling Laws https://arxiv.org/abs/2410.18837
- https://arxiv.org/pdf/2411.00247
- Generalization in diffusion models arises from geometry-adaptive harmonic representations https://arxiv.org/abs/2310.02557
- Arous G B, Gheissari R, Jagannath A. Online stochastic gradient descent on non-convex losses from high-dimensional inference. JMLR 2021.
- Damian A, Pillaud-Vivien L, Lee J, et al. Computational-statistical gaps in gaussian single-index models COLT 2024.