From Graphical Models to Generative AI: Mathematical Foundations, Algorithms, and Modern Applications

Michael Chertkov

Description: This course is designed for graduate students across a wide range of disciplines — including mathematical and computer sciences, engineering, and the physical, social, and biological sciences — seeking rigorous yet accessible mathematical foundations for modern AI. The course traces a coherent intellectual arc from classical probabilistic graphical models through the contemporary revolution in Generative AI, covering diffusion models, normalizing flows, variational autoencoders, transformers, and large language models (LLMs). Students will understand *why* generative AI works by grounding it in the principled mathematics of probabilistic inference, optimization, and information theory. The course draws heavily on the instructor's forthcoming textbook *Mathematics of Generative AI* (World Scientific, 2026) and accompanying open-source Python/Jupyter notebooks, equipping students with both theoretical depth and hands-on computational skills to tackle problems at the frontier of science and engineering.

Prerequisites: Familiarity with fundamental undergraduate mathematics (linear algebra, probability, analysis, and differential equations) is assumed. At least one graduate-level course in Applied Mathematics, Probability Theory, Statistics, Statistical Mechanics, or Machine Learning is strongly recommended. Some prior exposure to Python or another scientific computing language is helpful but not strictly required.

Assignments & Credits: Course evaluation includes a theory test in October and an individual project, with no homework, midterms, or finals. Grading is based on lecture attendance (10%), the theory test (40%), and the project presentation (50%). Over 30 project options will be provided alongside relevant research papers, and students are welcome to propose their own projects (subject to approval by September 15). Projects may be theoretical or include a software/implementation component; any modern scientific software (e.g., Python/PyTorch, Julia, MATLAB) is acceptable. Project presentations are planned for November–December.

Textbooks: The primary texts are the instructor's own manuscripts:

A. *Mathematics of Generative AI*, M. Chertkov, World Scientific (to appear, 2026). Living draft and Jupyter/Python notebooks available at https://github.com/mchertkov/Mathematics-of-Generative-AI-Book.
B. *Inference Learning and Optimization with Graphical Models*, M. Chertkov, https://sites.google. com/site/mchertkov/living-books/graphical-model-book

Additional recommended books and online resources include:

1. Graphical Models, Exponential Families, and Variational Inference, M. Wainwright & M. Jordan, Foundations and Trends in Machine Learning, 2008, ISBN-13: 978-1601981844.
2. High-Dimensional Statistics: A Non-Asymptotic Viewpoint, M. J. Wainwright, Cambridge University Press, 2019, ISBN-13: 978-1108498029.
3. Probabilistic Graphical Models: Principles and Techniques, D. Koller & N. Friedman, MIT Press, 2008, ISBN-13: 978-0262013192.
4. Information, Physics and Computation, M. Mézard & A. Montanari, Oxford University Press, 2009, ISBN-13: 978-0198570837.

5. Deep Learning, I. Goodfellow, Y. Bengio & A. Courville, MIT Press, 2016, ISBN-13: 978-0262035613. (Freely available at https://www.deeplearningbook.org)

6. Understanding Deep Learning, S. J. D. Prince, MIT Press, 2023, ISBN-13: 978-0262048644. (Freely available at https://udlbook.github.io/udlbook)

7. Reinforc0ement Learning: An Introduction, R. S. Sutton & A. G. Barto, MIT Press, 2018, ISBN-13: 978-02620392046.

Expected learning outcomes: After completion of the course, students will

- Understand the mathematical foundations shared by classical graphical models and modern generative AI architectures.
- Know how to formulate and solve structured probabilistic inference and learning problems using both exact and approximate methods.
- Understand the operational principles of modern generative models — including diffusion models, nor-malizing flows, VAEs, GANs, and autoregressive models — and their connections to classical inference.
- Appreciate the role of transformers and attention mechanisms as a new paradigm for representing and learning over structured data.
- Acquire research literacy and presentation skills (oral and written) sufficient to engage with the current literature.

Topic	**Summary of Topic**
Graphical Models — Setting the Stage	Structured statistical inference: sampling, marginal probabilities, partition functions, and free energies. Connections to physics (Ising models, spin glasses), information theory, and computer science. Motivating examples drawn from modern AI.
Computational Complexity & Algorithms	Deterministic and stochastic approaches. Statistical inference as optimization: Kullback–Leibler divergence, free energy functionals, variational principles. i.i.d. sampling and concentration phenomena.
Approximate Inference Methods	Variational approaches: Mean-Field, Belief Propagation, Linear Programming relaxations, Loop Series and cumulant expansions. Vari-able elimination and tensor-network methods. MCMC and FPRAS schemes. Interplay and synthesis of methods.
Latent Variable Models & Foundations of Generative Modeling	Expectation-Maximization. Restricted Boltzmann Machines and energy-based models as the historical precursors to deep generative AI. The Evidence Lower Bound (ELBO). Variational Autoencoders (VAEs): encoder–decoder architectures, reparameterization trick, la-tent space geometry. Connections between classical graphical model learning and modern deep latent-variable frameworks.
Normalizing Flows & Score-Based Models	Change-of-variables formula and exact likelihood. Planar, radial, and continuous normalizing flows; neural ODEs. Score matching and denoising score matching as a bridge to diffusion. Connections to Fokker–Planck equations and non-equilibrium statistical mechanics.

Topic	Summary of Topic
Diffusion Models & Stochastic Generative Processes	Forward noising processes and reverse-time SDEs (DDPM, DDIM (score-based diffusion). Mathematical foundations: Itô calculus, Langevin dynamics, and the reverse diffusion formula. Guidance, conditioning, and sampling acceleration. State-of-the-art image, audio, and scientific (molecular, climate) applications.
Autoregressive Models & Attention	Autoregressive factorization as a directed graphical model. Sequence models: RNNs, LSTMs, and their limitations. The Transformer: self-attention as a new inference primitive, positional encodings, multi-head attention. Large Language Models (LLMs): scaling laws, in-context learning, emergent capabilities, and open problems.
Learning, Fine-Tuning & Alignment	Learning of generative models as (often non-convex) optimization: stochastic gradient descent, Adam, and modern variants. Pre-training, fine-tuning, and parameter-efficient methods (LoRA, adapters). Reinforcement Learning from Human Feedback (RLHF) as graphical-model inference. Constitutional AI and alignment challenges.
Generative AI for Science & Engineering	Depending on student interest, additional theoretical and application topics will be covered. Possible theoretical topics: (a) Optimal Transport and the Wasserstein perspective on generative models; (b) Gaussian graphical models and linear diffusion; (c) Generative models on graphs (GNNs, graph diffusion). Possible application topics: (a) Protein structure and molecular design; (b) Climate and fluid dy-namics emulation; (c) Power and energy systems; (d) Scientific foundation models. Open problems and research frontiers.