Fall 2023

STAT 221: Computational Tools for Statistical Learning: Approximation, Optimization, and Monte Carlo

Instructor: Jun Liu; Email: jliu@stat.harvard.edu;
Teaching Fellow: Xiaodong Yang; Email: xiaodong_yang@fas.harvard.edu

COURSE DESCRIPTION: Focus on computational tools for statistical inference and learning. It differs from other courses on computational methods in its emphases on both computational thinking and statistical implications of inference algorithms. Topics include: basic algebraic operations; optimization methods; the EM and variational approximations; Markov chain Monte Carlo; importance sampling; sequential Monte Carlo; stochastic approximation; random forests; neural networks.

PREREQUISITE & REQUIREMENT: Advanced linear algebra and multivariate calculus (MATH 21a/b or equivalents); STAT 110 and 111 or equivalents. Knowledge of programming, especially R, and/or Python, and/or Matlab (or C/C++). Required to implement some of the methods discussed in class as part of the problem sets, and some amount of programming in the final project.

Class Meetings: **MW 3:00-4:15 PM**; Location: **Science Ctr 706**.
Office Hours: Monday 10:30-11:50 in my office (SC 715), **Tuesday** 8:30-9:30 PM (Eastern Time) via Zoom: https://harvard.zoom.us/my/junsliu, Passcode: 12345; or by **appointment**.

Policies for Homework, Participation, Generative AI, and Grading:

- **Grading:** The course grade is a weighted average of class participation (10%), homework (70%), and the final project (20%). The course is letter-graded by default, but may be switched to SAT/UNSAT at your request, in which case please inform me ASAP.
- **Participation:** Active participation is highly valued, including attending lectures and sections regularly, raising questions during lectures and sections, challenging instructors (intellectually), and engaging in other forms of discussions.
- You can discuss homework problems with other fellow classmates, but you must write up your own solutions in your own words. Additionally, you should list the names of the students with whom you collaborated (if any).
- Copying someone elses solution, or just making trivial changes for the sake of not copying verbatim, is not acceptable.
- You may use ChatGPT or other similar algorithms to help your study and to brainstorm, but beware that their generated responses may be inaccurate, incomplete, or mistaken.
- You may not submit any work generated by an AI program as your own. If you include material generated by an AI program, it should be cited like any other reference material.

- Your homework must be submitted as a correctly rotated (i.e., not upside-down or sideways) single PDF file via the Canvas course website; no paper submissions. Homework can be typeset, written using a tablet, or scanned from handwritten (clear and easily legible) notes.
- You may be asked to submit your computer codes, which need to be in the language of either R or Python or in an executable form with clear instructions.

<u>TOPICS TO BE COVERED</u> (not in chronological order and subject to changes):

- Classic Numerical Methods
  - Mode finding: gradient descent, conjugate gradient, Newton-Raphson, etc.
  - Numerical integration
  - Linear programming, dynamic programming (recursion)
  - Matrix operations: QR, Cholesky, SVD, PCA, etc.
- Monte Carlo Methods
  - Acceptance-Rejection, Antithetic variates, Control variates
  - Importance Sampling, sequential Importance Sampling, sequential Monte Carlo
  - Markov chain Monte Carlo, bootstrap
  - Hamiltonian dynamcis and hybrid Monte Carlo
  - Examples and other advanced MC techniques: collapsing, grouping, multiple-try Metropolis, bridge sampling, Wang-Landau algorithm, etc.
- Modern Tools for Statistical Learning
  - The EM algorithm, and its variations
  - Variational approximations, and others
  - Tree-based methods (CART, random forests, etc.)
  - Other topics (optional): boosting, neural networks (e.g., FNN, RNN, CNN), stochastic gradient descend (SGD), Variational Auto Encoder (VAE), FDR controls, etc.

<u>REFERENCES</u>

Fischetti, T. (2015). *Data Analysis with R*, Packt Publishing, ISBN-10: 1785288148.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning (Adaptive Computation and Machine Learning).* MIT Press. Freely available at `http://www.deeplearningbook.org/`.

Liu, J.S. (2001). *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag: New York. ISBN-10: 0387763694.

Nielson, M (2017). *Neural Networks and Deep Learning.* <u>Free on-line book</u>.

Thisted, R. (1988). *Elements of Statistical Computing.* Chapman and Hall/CRC, ISBN-10: 0412013711.

Wickham , H. and Garrett Grolemund (2018). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (1st Edition). O'Reilly Media. ISBN-10: 1491910399