HARVARD T.H. CHAN | SCHOOL OF PUBLIC HEALTH

BST234 Introduction to Data Structures and Algorithms

Spring 2023
Monday and Wednesday, 11:30 AM - 1:00 PM
Kresge 201

Instructor Information
Georg Hahn
655 Huntington Avenue
Building 1, 4th Floor, Room 435
Boston, MA 02115
Email: ghahn@hsph.harvard.edu
Office Hours: Building I, 4th Floor, Room 435, time to be confirmed after poll in class

Teaching Assistants
Max Wang, maxwang@g.harvard.edu

Credits
5 credits

Course Purpose and Description
Introduction to important computational problems in biostatistics and state-of-the-art algorithms for solving them. The course focuses on the implementation of data structures and algorithms to solve problems of practical relevance.
The courses BST221 and BST234 are based on the same material. The two courses differ in that BST234 (the PhD version) will have a more theoretical focus in class and in the assignments, and BST221 (the Master's version) will have a more applied focus in class and in the assignments.

Pre-Requisites
There are no pre-requisites for this course.

Course Learning Objectives
Upon successful completion of this course, you should be able to:

- Work with complex data structures and algorithms in practical settings, obtain a first exposure to proving the correctness of an algorithm, as well as its runtime analysis.
- Apply state-of-the-art numeric algorithms to efficiently solve linear equations, least squares problems, and eigenvalue/eigenvector computations which are at the heart of countless practical problems in (bio-)statistics and physical sciences.
- Have a detailed understanding of a variety of important computational problems and the state-of-the-art algorithms for solving them. Train the ability to identify what (known) core problem lays at the heart of a problem encountered in practice.
- Integrate functions numerically and with Monte Carlo methods.
- Have a broad understanding of the difficulties arising with non-linear equations and general purpose functional optimization and apply state-of-the-art algorithms to solve such problems.
- Create implementations of algorithms in R or Python, and apply them to solve real-world problems.
- Develop your own algorithmic solutions for biostatistical problems.

Course Readings
1. “Introduction to Algorithms” by Cormen, Leiserson, Rivest, Stein; Second Edition; MIT Press.
2. “Fundamentals of Matrix Computations” by Watkins; Third Edition; Wiley.

Course Structure
Classroom interactions rely on attendance of students,, their in-class participation, and their mutual respect for each other. All technical aspects of the course (time and place of the lectures, lecture slides, exercise sheets and zoom recordings) will be handled in Canvas. The lecture hours are used to deliver the lecture and engage the students in interactive activities/ discussions to illustrate the theoretical concepts. Recordings are available after the lecture on Canvas. The labs are used to discuss the assignments. Office hours are provided by both the instructor of the course as well as by the teaching assistants.

The final grade for this course will be based on all 5 graded homework assignments, each counting 1/5 towards ½ of the final grade. The other half will be covered by midterm and final (equal amounts).

Required format: All homework assignments must be 12 point font or larger, single spaced, and 1” margins. The assignments need to be submitted online on canvas within the time window specified on each homework. Late assignments are usually not accepted. Each student must individually write their own answers to the homework assignments. They may, and are encouraged to, work together in groups to discuss the homework readings.

Participation
Discussions in class and active participation in class are a vital part of the learning process, and all students are expected to attend and participate in all classes. Participation is not graded.

Homeworks (50%)
All graded homeworks together equally count towards 50% of the final grade. All graded assignments are posted on Canvas. The grading is done with respect to correctness of your answer. The grading scheme (number of points per question) is given on each homework assignment. Late assignments are not accepted. For any special arrangements (medical reasons etc.) please contact the instructor.

Midterm (25%)
The midterm counts towards 25% of the final grade. The grading is done with respect to correctness of your answer. The grading scheme (number of points per question) is given on the midterm. Late assignments are not accepted. For any special arrangements (medical reasons etc.) please contact the instructor.

Final (25%)
The final counts towards 25% of the final grade. The grading is done with respect to correctness of your answer. The grading scheme (number of points per question) is given on the final. Late assignments are not accepted. For any special arrangements (medical reasons etc.) please contact the instructor.

Additional Information
All components – lectures, seminars/laboratories, and small group sessions – are integral and mandatory parts of the course. You are expected to read any materials provided prior to the class session and are strongly encouraged to attend and be prepared for class discussions. If you are unable to attend the class session due to time zone differences, we encourage you to work through the material on your own and watch the recorded discussions. Optional additional class sessions may be offered by faculty and teaching assistants at other times to accommodate different time zones.

Technical Information
Assistance

Canvas

If the issue is Canvas-related (e.g., you can't figure out how to use something or a feature seems broken), first try the documentation located under the Help menu found on the left-hand side of each Canvas page. If the issue is not covered there, contact Instructure directly, also via the Help menu. You can e-mail, text, or speak live with them at any time day or night. If you cannot access Canvas to view the Help menu, you can reach Instructure by phone at +1 (844) 326-4466.

Zoom

For help with Zoom video conferencing, first check the variety of video tutorials and online help at https://support.zoom.us. In addition, you may contact the Helpdesk by emailing helpdesk@hsph.harvard.edu or calling +1 (617) 432-HELP (4357).

Harvard-Specific Issues

If the issue seems Harvard-specific (e.g., HUID or myHarvardChan authentication, email not working, etc.), contact the Helpdesk at helpdesk@hsph.harvard.edu or +1 (617) 432-HELP (4357).

Other

If you are unsure where to turn, but think the issue is related to technology or the course lecture videos, contact the Helpdesk as noted above.

Technical Requirements

- Reliable, high-speed internet connection
- Your laptop must meet the minimum technical requirements found on the Student Guide page.
- Modern and updated web browser (e.g., a recent version of Firefox or Chrome)
- Web camera and microphone (integrated into computer or USB peripheral)
- Throughout this program, you will be using VDI to access certain applications (e.g., EndNote and JMP); in turn, your computer must meet the minimum hardware and software requirements displayed on the VDI page.
- Please contact helpdesk@hsph.harvard.edu with questions.

Please note that while it is possible to access most of the course materials via mobile and wireless devices, video conferencing and other bandwidth-intensive sessions will have the greatest reliability on a wired high-speed connection.

Harvard Chan Policies and Expectations
Inclusivity Statement

Diversity and inclusiveness are fundamental to public health education and practice. Students are encouraged to have an open mind and respect differences of all kinds. I share responsibility with you for creating a learning climate that is hospitable to all perspectives and cultures; please contact me if you have any concerns or suggestions.

Bias Related Incident Reporting

The Harvard Chan School believes all members of our community should be able to study and work in an environment where they feel safe and respected. As a mechanism to promote an inclusive community, we have created an anonymous bias-related incident reporting system. If you have experienced bias, please submit a report here so that the administration can track and address concerns as they arise and to better support members of the Harvard Chan community.

Title IX
The following policy applies to all Harvard University students, faculty, staff, appointees, or third parties: Harvard University Sexual and Gender-Based Harassment Policy. Procedures For Complaints Against a Faculty Member
Procedures For Complaints Against Non-Faculty Academic Appointees

Academic Integrity
Each student in this course is expected to abide by the Harvard University and the Harvard T.H. Chan School of Public Health School's standards of Academic Integrity. All work submitted to meet course requirements is expected to be a student's own work. In the preparation of work submitted to meet course requirements, students should always take great care to distinguish their own ideas and knowledge from information derived from sources.

Students must assume that collaboration in the completion of assignments is prohibited unless explicitly specified. Students must acknowledge any collaboration and its extent in all submitted work. This requirement applies to collaboration on editing as well as collaboration on substance.

Should academic misconduct occur, the student(s) may be subject to disciplinary action as outlined in the Student Handbook. See the Student Handbook for additional policies related to academic integrity and disciplinary actions.

Accommodations for Students with Disabilities
Harvard University provides academic accommodations to students with disabilities. Any requests for academic accommodations should ideally be made before the first week of the semester, except for unusual circumstances, so arrangements can be made. Students must register with the Local Disability Coordinator in the Office for Student Affairs to verify their eligibility for appropriate accommodations. Contact Colleen Cronin ccronin@hsph.harvard.edu in all cases, including temporary disabilities.

Religious Holidays, Absence Due to
According to Chapter 151c, Section 2B, of the General Laws of Massachusetts, any student in an educational or vocational training institution, other than a religious or denominational training institution, who is unable, because of his or her religious beliefs, to attend classes or to participate in any examination, study, or work requirement on a particular day shall be excused from any such examination or requirement which he or she may have missed because of such absence on any particular day, provided that such makeup examination or work shall not create an unreasonable burden upon the School. See the student handbook for more information.

Grade of Absence from Examination
A student who cannot attend a regularly scheduled examination must request permission for an alternate examination from the instructor in advance of the examination. See the student handbook for more information.

Final Examination Policy
No student should be required to take more than two examinations during any one day of finals week. Students who have more than two examinations scheduled during a particular day during the final examination period may take their class schedules to the director for student affairs for assistance in arranging for an alternate time for all exams in excess of two. Please refer to the student handbook for the policy.

Course Evaluations

Constructive feedback from students is a valuable resource for improving teaching. The feedback should be specific, focused and respectful. It should also address aspects of the course and teaching that are positive as well as those which need improvement.

Completion of the evaluation is a requirement for each course. Your grade will not be available until you submit the evaluation. In addition, registration for future terms will be blocked until you have completed evaluations for courses in prior terms.

Course Schedule

Objectives	Readings	Assignments/Activities
Week 1		
Session 1. Welcome to Course		
Upon successful completion of this session, you should be able to: • Outline key aspects of the course as explained in the syllabus • Identify course policies • Navigate through the course site	N/A	No problem set.
Session 2. Random number generation		
Upon successful completion of this session, you should be able to: • Know about pseudo-random number generators • Know desired properties and drawbacks • Know current gold standard	Entry "Random number generation" on Wikipedia: https://en.wikipedia.org/wiki/Random_number_generation	
Week 2		
Session 3. Basic concepts of algorithms and complexity		
Upon successful completion of this session, you should be able to: • Know about concepts of algorithms: What is an algorithm? What flavors of algorithms exist? • Identify algorithmic complexity in time and space	Chapter I of Cormen et al.	Problem set 1 on random number generation.
Session 4. Application to sorting algorithms		

Week 2		
Upon successful completion of this session, you should be able to: • Know the current state-of-the-art sorting algorithms • Apply recursion and divide-and-conquer principles to design a sorting algorithm	Chapter II of Cormen et al.	

Week 3		
Session 5. Examples of basic data structures		
Upon successful completion of this session, you should be able to: • Know about different types and benefits of popular data structures • Handle array/ vector, list, (doubly) linked list, stack, queue • Follow implementation examples	Section III.10 of Cormen et al.	Problem set 2 on Big-O notation and runtime.
Session 6. Examples of basic data structures (continued)		
Upon successful completion of this session, you should be able to: • Know about trees: binary, n-ary, terminology, traversal • Handle binary search trees and operations on such trees • Know about heaps and heapsort	Section III.12 of Cormen et al.	

Week 4		
Session 7. Heaps		
Upon successful completion of this session, you should be able to: • Understand the heap data structure and know how to execute operations on it.	Section II.6 of Cormen et al.	Problem set 3 on recursion as well as algorithm design and correctness.
Session 8. Heaps (continued)		
Upon successful completion of this session, you should be able to: • Know how the heap data structure can be used to design an efficient sorting algorithm (heap sort)	Section II.6 of Cormen et al.	

Week 5		
Session 9. Greedy algorithms		
Upon successful completion of this session, you should be able to: • Apply the divide-and-conquer concept • Understand the idea behind greedy algorithms, in particular the greedy choice property and optimal substructure principle	Section IV.16 of Cormen et al.	Problem set 4 on tree traversals and recursive sorting algorithms.
Session 10. Greedy algorithms (continued)		
Upon successful completion of this session, you should be able to: • Identify the Knapsack problem and NP-completeness • Know about dynamic programming	Section IV.16 of Cormen et al.	

Week 6		
Session 11. Parallel programming		
Upon successful completion of this session, you should be able to: • Define speedup, embarrassingly parallel computations, parallel architectures, synchronization and locks • Know about message passing (MPI), Map/Reduce, and know software tools	Entry "Parallel computing" on Wikipedia: https://en.wikipedia.org/wiki/Parallel_computing	Problem set 5 on Quicksort and algorithm design.
Session 12. P and NP		
Upon successful completion of this session, you should be able to: • Identify complexity classes • Define NP-hard and NP-complete problems, know classic examples thereof and current state-of-the-art solving techniques	Entry "P versus NP problem" on Wikipedia: https://en.wikipedia.org/wiki/P_versus_NP_problem	

Week 7		
Session 13. Numerical aspects of computer algorithms		
Upon successful completion of this session, you should be able to: • Know the binary representation of floating point numbers in a computer, identify precision and accuracy • Know vector norms, relative and absolute numerical errors	Chapter 2 of Watkins	Problem set 6 on Kruskal's algorithm and a greedy solution of the fractional knapsack problem.
Session 14. Numerical aspects of computer algorithms (continued)		
Upon successful completion of this session,	Chapter 2 of Watkins	

Week 7		
you should be able to: • Identify well-posed and ill-posed numerical problems, define the condition number • Identify sources of numerical errors • Assess the stability of linear equation systems		

Week 8		
Session 15+16. Midterm		
This week is reserved for the take-home midterm exam.		

Week 9		
Session 17. Systems of linear equations		
Upon successful completion of this session, you should be able to: • Identify singular and nonsingular matrices • Solve linear equation systems, especially triangular systems	Chapter 1 of Watkins	Take-home midterm exam.
Session 18. Systems of linear equations (continued)		
Upon successful completion of this session, you should be able to: • Perform Gauss elimination, LU and Cholesky decompositions	Chapter 1 of Watkins	

Week 10		
Session 19. Least squares problem and numerical solutions		
Upon successful completion of this session, you should be able to: • Define and apply the least squares method and know several solution approaches • Compute the QR decomposition	Chapter 3 of Watkins	Problem set 7 on Dijkstra's algorithm and variants of the shortest path problem.
Session 20. Least squares problem and numerical solutions (continued)		
Upon successful completion of this session, you should be able to: • Define and perform Givens rotations and Householder transformations	Chapter 3 of Watkins	

Week 11		
Session 21. Eigenvalues and eigenvectors		
Upon successful completion of this session, you should be able to: • Define eigenvalues and know popular areas of application • Know about numerical issues arising with the computation of eigenvalues	Chapter 5 of Watkins	Problem set 8 on deriving an approximation algorithm for the NP-complete "bin packing" problem.
Session 22. Systems of nonlinear equations		
Upon successful completion of this session, you should be able to: • Know basic properties of nonlinear equations, sensitivity and conditioning, convergence rate • Know state-of-the-art methods for solving them	Chapter 5 of Watkins	

Week 11		
Week 12		
Session 23. Numerical integration and MCMC		
Upon successful completion of this session, you should be able to: • Know about permutation testing, null distribution, and p-value approximation	Entry “Monte Carlo integration” on Wikipedia: https://en.wikipedia.org/wiki/Monte_Carlo_integration	Problem set 9 on computing the condition number, LU and Cholesky decompositions.
Session 24. Numerical integration and MCMC (continued)		
Upon successful completion of this session, you should be able to: • Apply numerical integration and perform Monte Carlo integration • Know how importance sampling and the Metropolis-Hastings algorithm work • Define MCMC	Entries “Markov chain Monte Carlo” and “Metropolis-Hastings algorithm” on Wikipedia: https://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo https://en.wikipedia.org/wiki/Metropolis%E2%80%93Hastings_algorithm	

Week 13		
Session 25. Numerical optimization		
Upon successful completion of this session, you should be able to: • Apply different types of optimization methods • Identify unconstrained and constrained optimization • Define the Karush-Kuhn-Tucker conditions • Apply line search and steepest descent • Know about issues arising with saddle points, valleys, collinearity	Entry “Karush-Kuhn-Tucker conditions” on Wikipedia: https://en.wikipedia.org/wiki/Karush%E2%80%93Kuhn%E2%80%93Tucker_conditions	Problem set 10 on Monte Carlo integration.

Week 13		
Session 26. Numerical optimization (continued)		
Upon successful completion of this session, you should be able to: • Apply ridge regression • Carry out Newton-Raphson in higher dimensions • Know about quasi-Newton methods • Define the conjugate gradient method • Apply probabilistic optimization, especially simulated annealing	Entries “Quasi-Newton method” and “Simulated annealing” on Wikipedia: https://en.wikipedia.org/wiki/Quasi-Newton_method https://en.wikipedia.org/wiki/Simulated_annealing	

Week 14		
Session 27+28. Final project		
This week is reserved for the final group project. The students form groups and select from a range of topics.		Problem set 11 on stochastic gradient descent.

Week 15		
Session 29+30. Class presentations		
This week is reserved for in-class presentations of all groups that worked on the final project.		