Oishi Banerjee and Cooper Raterink
CS109

Handout #1
Updated June 16, 2020

Administrivia

Instructors: Oishi Banerjee and Cooper Raterink
Oishi's Email: `oishib@stanford.edu`
Cooper's Email: `crat@stanford.edu`

Class web page: `http://cs109.stanford.edu`
CS109 staff email: `cs109-sum1920-staff@mailman.stanford.edu`

You should regularly check the class web site for the office hours schedule, class announcements and other information, including the most up-to-date details on assignments and errata. The class web page will also have the schedule of topics to be covered and links to other class materials, including class handouts.

The staff email address reaches the whole CS109 teaching staff in case you have personal questions. All other questions about course content should be posted on our Ed discussion forum.

Prerequisites

The prerequisites for this course are CS103, CS106B or X, and Math 51 (or equivalent courses). Probability involves a fair bit of mathematics (set theory, calculus and familiarity with linear algebra), and we'll be considering several applications of probability in computer science that require familiarity with algorithms and data structures covered in CS106B/X. Since some students who may not have taken CS103 and/or Math 51 at Stanford have asked what level of mathematics background is necessary, here is a quick rundown of some of the mathematical tools we'll be using in this class: calculus (integration and differentiation), linear algebra (basic operations on vectors and matrices), an understanding of the basics of set theory (subsets, complements, unions, intersections, cardinality, etc.), and familiarity with basic proof techniques (including induction). If you're not already comfortable with these mathematical topics, please be warned that CS109 may prove especially difficult for you.

What are CS109 and Probability Theory all about?

While the initial foundations of computer science began in the world of discrete mathematics (after all, modern computers are digital in nature), recent years have seen a surge in the use of probability as a tool for the analysis and development of new algorithms and systems. As a result, it is becoming increasingly important for budding computer scientists to understand probability theory, both to provide new perspectives on existing ideas and to help further advance the field in new ways.

Probability is used in a number of contexts, including analyzing the likelihood that various events will happen, better understanding the performance of algorithms (which are increasingly making use of randomness), or modeling the behavior of systems that exist in asynchronous environments ruled by uncertainty (such as requests being made to a web server). Probability provides a rich

set of tools for modeling such phenomena and allowing for precise mathematical statements to be made about the performance of an algorithm or a system in such situations.

Furthermore, computers are increasingly often being used as data analysis tools to glean insights from the enormous amounts of data being gathered in a variety of fields; you've no doubt heard the phrase "big data" referring to this phenomenon. Probability theory is now used as one of the primary methods for designing new algorithms to model such data, allowing, for example, a computer to make predictions about new or uncertain events. In fact, many of you have already been the users of such techniques. For example, most email systems now employ automated spam detection and filtering. Methods for being able to automatically infer whether or not an email message is spam are frequently rooted in probabilistic methods. Similarly, if you have ever seen online product recommendation (e.g., "customers who bought X are also likely to buy Y"), you've seen yet another application of probability in computer science. Even more subtly, answering detailed questions like how many buckets you should have in your a hash table or how many machines you should deploy in a data center (server farm) for an online application make use of probabilistic techniques to give precise formulations based on testable assumptions.

Our goal in this course is to build foundational skills and give you experience in the following areas:

1. Understanding the combinatorial nature of problems: Many real problems are based on understanding the multitude of possible outcomes that may occur, and determining which of those outcomes satisfy some criteria we care about. Such understanding is important both for determining how likely an outcome is, but also for understanding what factors may affect the outcome (and which of those may be in our control).

2. Working knowledge of probability theory: Having a solid knowledge of probability theory is essential for computer scientists today. Such knowledge includes theoretical fundamentals as well as an appreciation for how that theory can be successfully applied in practice. We hope to impart both these concepts in this class.

3. Appreciation for probabilistic statements: In the world around us, probabilistic statements are often made, but are easily misunderstood. For example, when a candidate in an election is said to have a 53% likelihood of winning does this mean that the candidate is likely to get 53% of the vote, or that that if 100 elections were held today, the candidate would win 53% of them? Understanding the difference between these statements requires an understanding of the model in the underlying probabilistic analysis.

4. Applications: We are not studying probability theory simply for the joy of drawing summation symbols (okay, maybe some people are, but that's not what we're really targeting in this class), but rather because there are a wide variety of applications where probability allows us to solve problems that might otherwise be out of reach (or would be solved more poorly without the tools that probability can bring to bear). We'll look at examples of such applications throughout the class.

5. An introduction to machine learning: Machine learning is a quickly growing subfield of artificial intelligence which has grown to impact many applications in computing. It focuses

on analyzing large quantities of data to build models that can then be harnessed in real problems, such as filtering email, improving web search, understanding computer system performance, predicting financial markets, or analyzing DNA.

The class starts by providing a fundamental grounding in combinatorics, and then quickly moves into the basics of probability theory. We will then cover many essential concepts in probability theory, including particular probability distributions, properties of probabilities, and mathematical tools for analyzing probabilities. Finally, the last third of the class will focus on data analysis and machine learning as a means for seeing direct applications of probability in this exciting and quickly growing subfield of computer science.

Goals for Summer 2020

We fully recognize that this experience cannot replace what we normally have on campus, and that many of you have personal situations that may change throughout the quarter. That being said, we are determined to reach the following course goals to the best of our ability: (1) To maintain the intellectual rigor of the CS109 curriculum while providing flexible ways for you to learn, and (2) To foster and maintain human connections and a sense of community throughout this online course. We have adjusted the typical course structure to meet these goals.

Many of these changes that we are making are designed to help foster intellectual nourishment, social connection, and personal accommodation—through accessible, asynchronous content for diverse access, time zones, and contexts, and optional, synchronous discussion to learn together and combat isolation. Please bear in mind that all of us have limited experience working with remote, online classrooms at this scale, and we may have to adapt throughout the quarter. Everyone needs support and understanding in this unprecedented moment, and we are here to listen to you. Thanks and welcome to CS109. (Credit for this wording goes to Prof. Brandon Bayne from UNC - Chapel Hill.)

Lecture

We will be holding live lectures remotely via Zoom on Mondays, Wednesdays, and Fridays, starting at 1:30PM Pacific Time. The live lecture will not use the full 110-minute slot listed on Axess. Lecture content will be covered in two parts:

Asynchronous recorded videos pre-lecture. We have pre-recorded roughly 30-60 minutes worth of content in short videos to cover introductory material each lecture. You should watch and be familiar with this material prior to attending live lecture. To help prepare you for effective participation in live lecture, we will have "concept checks," which are required self-tests on the pre-recorded content that will count towards your grade. To encourage you to stay on track, 15% of your course grade will come from completing concept checks.

Synchronous, live lecture. The live lecture is dedicated to interactive problem solving and advanced concept learning. We'd love for you to ask questions and work together in small groups. We realize that many of you have various constraints on your schedule due to timezone differences, other academic commitments, or personal circumstances, and so all live lectures will be recorded and posted shortly after lecture (there may be a delay of up to 24 hours). To encourage discussion even after lecture has ended, we will be using our Ed discussion forum (not Zoom chat) to manage questions during lecture.

Units and degree requirements

If you are an undergraduate, you are **required** to take CS109 for 5 units of credit. If you are a graduate student, you may enroll in CS109 for 3 or 4 units if it is necessary for you to reduce your units for administrative reasons. Taking the course for reduced units does not imply any change in the course requirements.

Handouts and Textbook

We have lecture notes for all of the lectures so that students can follow the class without the optional textbook, listed below.

You are responsible for material covered in the class lectures and course handouts. So it probably goes without saying that it would be a good idea to attend the live sessions if you are able. At a minimum, we recommend that you keep pace with the concept check schedule if your situation allows it.

The *optional* textbook for this course is:

Sheldon Ross, *A First Course in Probability* (10th Ed.), Pearson Prentice Hall, 2018.

The 8th, 9th, and 10th editions of the textbook are fine for this class. The text is not required material, but students may find Ross offers a different and useful perspective on the important concepts of the class. Suggested, optional reading assignments from the textbook (10th Ed.) are in the schedule on the course website.

HathiTrust, a library archive of which Stanford is a member, has granted the university online access to the 8th edition (2010) for the duration of the Summer quarter. The "check out" system works similarly to print reserves: A user can check out the book an hour at a time as long as they are actively using it. Access guidelines are on the HathiTrust How To Use It webpage (`https://www.hathitrust.org/ETAS-How-To`), and once you're logged in, the book is at this link: `https://catalog.hathitrust.org/Record/006904347`

All students should retain receipts for books and other course-related expenses, as these may be qualified educational expenses for tax purposes. If you are an undergraduate receiving financial aid, you may be eligible for additional financial aid for required books and course materials if these expenses exceed the aid amount in your award letter. For more information, review your award letter or visit the Student Budget website (`https://financialaid.stanford.edu/undergrad/budget/index.html`).

Assignments and Exams

During the course, there will be six problem sets assigned. **Each student is to submit individual work on the problem sets** (more details on that below when we talk about the Honor Code). All homework assignments should be turned in **30 minutes before the beginning of class** on the respective due date for each assignment.

In addition to the assignments, there will two quizzes, each weighted equally. The first is scheduled Monday, July 20th, and the second will take place during the end-quarter period.

Each quiz is a take-home, open-notes exam that approximates about 1-2 hours of active work and should be completed individually. To reduce time pressure and accommodate any difficulties accessing resources, you will have at least 24 hours to work on each quiz. If you have an exceptional circumstance regarding either of these dates and absolutely cannot make any part of the 24-hour exam window, you must let us know via the Google Form we send out in the first week.

As mentioned above, on the assignments and exams you will be responsible for material covered in the class lectures and course handouts.

Discussion Sections. Each week for 50 minutes you have the option to meet with one of the instructors to work through problems. Section will be recorded. If you have taken any of the CS106 classes, our sections will be very similar—except with more probability. While section attendance is not graded, we encourage you to attend, listen to the recording or work through the section problems on your own.

Grading

The grade for the course will be determined according to the following breakdown:

Assignments ($6 \times 10\%$):	60%
Quizzes ($2 \times 12.5\%$):	25%
Participation:	15%

The participation component of the grade consists mainly of concept checks, short quizzes submitted to Gradescope before each lecture. We reserve the right to increase your participation grade based on active participation on the Ed discussion board, in class or in sections.

You are guaranteed a minimum letter grade according to standard grade cut-offs. For example, if your final grade is 94% or higher, you are guaranteed to get at least an A. At the end of the term, we may assign even higher letter grades. In other words, any curve we apply can only help you.

Late Assignments

We anticipate more than ever that during this quarter, there may be unforeseen circumstances that make it difficult to turn in homework assignments on time.

For concept checks, there will be no deductions for lateness. Concept checks may be submitted until August 13th at 11:59pm PDT.

For homework assignments, we will use the following late policy: If you submit an assignment by the on-time deadline listed on the assignment writeup, you receive an extra credit on-time bonus, typically 5%. If you submit by up to 1 class day late, you do not receive this bonus; up to 2 class days late, your submission is capped at 80% of the maximum points; up to 3 class days late, capped at 60% maximum points. Note that no assignment will be accepted more than a week late, since we will be releasing solutions to the assignments one week after their due date.

Each late day counts as a **class** day. Therefore, a homework assignment due on Monday may be turned on Wednesday before 10:00am and count as one day late. A homework assignment due on Friday that is turned in on the following Monday is also counted as one day late.

As an example, suppose a student was set to receive 95% on an assignment due Friday. If turned in on-time, they receive a score of 100%; if turned in by the following Monday, they receive 95%; if turned in by the following Wednesday, they receive 80%; if turned in by the following Friday, a week after the original deadline, they receive 60%.

Note that no assignments are accepted beyond the last class day of the regular quarter (Thursday, August 13th). If you have a personal or medical emergency, any extensions must be cleared prior to the assignment being due, and the request should be made directly to Cooper and Oishi.

Honor Code

Each student is expected to do their own work on the problem sets in CS109. Students may discuss problem sets with each other as well as the course staff. **Any discussion of problem set questions with others should be noted on a student's final write-up of the problem set answers.** Each student must turn in their own write-up of the problem set solutions. Excessive collaboration (i.e., beyond discussing problem set questions) can result in honor code violations. Questions regarding acceptable collaboration should be directed to the class instructor *prior* to the collaboration. **It is a violation of the honor code to copy problem set or exam question solutions from others, or to copy or derive them from solutions found online or in textbooks, previous instances of this course, or other courses covering the same topics (e.g., STATS 116 or probability courses at other schools).** Copying of solutions from students who previously took this or a similar course is also a violation of the honor code. Finally, a good point to keep in mind is that you must be able to explain and/or re-derive anything that you submit.

Students with Documented Disabilities

Students who may need an academic accommodation based on the impact of a disability must initiate the request with the Office of Accessible Education (OAE). Professional staff will evaluate the request with required documentation, recommend reasonable accommodations, and prepare an Accommodation Letter for faculty. For students who have disabilities that don't typically change appreciably over time, the letter from the OAE will be for the entire academic year; other letters will be for the current quarter only. Students should contact the OAE as soon as possible since timely notice (for example, at least a week before an exam) is needed to coordinate accommodations. Students should also send your accommodation letter to instructors as soon as possible. **If you require additional, or different, accommodations specific to the Summer 2020 learning environment, please contact your disability adviser directly.**