David Varodayan
CS109

Handout #1
January 6, 2020

Administrivia

Instructor:	David Varodayan
Email:	varodayan@stanford.edu
Office:	Gates 161
Office Hours:	Wednesdays 3:00-5:00pm in Gates 161 and by appointment
Class web page:	http://cs109.stanford.edu
CS109 staff email:	cs109@cs.stanford.edu

You should regularly check the class web site for announcements and other information, including the most up-to-date information 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 specific questions about the course material, assignments, or lectures.

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). We'll also do combinatorics in the class, but we'll be covering a fair bit of that material ourselves in the first week. If you don't feel comfortable with these mathematical topics and still insist on taking CS109 without having taken CS103 and Math 51, we sincerely regret the pain that this class may cause you—but don't say you weren't warned. I mean, after all, we just warned 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.

"Working" Office Hours

To help make you more successful in this class , the course staff will hold "working" office hours (aside from David's office hours, which are primarily for conceptual and course-level questions). The idea is to encourage you to work on your problem sets at these office hours, so you can immediately ask any questions that come up while working on your problem sets. While you are certainly not required to attend any of these working office hours, they are simply meant to encourage you to interact with the course staff more often in order to help you better understand the course material. Besides, our job is to help everyone learn the material for this class, and being more accessible to you when you are actually working on your assignments (rather than when you just have a problem) will help the course go more smoothly for you (and it'll be more fun for us).

Units

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

The *optional* textbook for this course is:

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

Several copies of the textbook will be held on reserve in the Engineering Library and Science Library. The Stanford Bookstore should also have some copies in stock, or you can order it online from your favorite book shopping site on the web. The 8th or 9th Editions of the textbook are also fine for this class. We have lecture notes for all of the lectures so that students can follow the class without the Ross textbook. 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 are in the schedule on the course website.

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`).

You are responsible for material covered in the class lectures and course handouts, not the readings in the textbook (which are supplementary). So it probably goes without saying that it would be a *really* good idea to go to class, as it's likely to increase your probability of getting a good grade. :-)

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 be a 2-hour midterm exam and a 3-hour final exam. The midterm exam will be administered on **Monday, February 10, from 7:00pm to 9:00pm**. If you have an exceptional circumstance regarding the midterm and absolutely cannot make the scheduled midterm exam, you must let us know via the Google Form on question 1 of the first problem set.

The final examination is scheduled for **Wednesday, March 18, from 3:30pm-6:30pm**. For a variety of reasons (including university policy), there will be **no alternate time for the final exam**. Please make sure that you can attend the final exam at the specified time before enrolling in the class. If you cannot make the final exam, you should plan on taking CS109 in a different quarter.

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

Participation

Active participation plays an important role in making you a master of probability for computer science. It has also been observed over many quarters that keeping up with the material highly correlates with improved class performance. To encourage you to stay on track, 5% of your course grade will come from attending and actively participating in discussion sections.

Discussion Sections. Each week for 50 minutes you will meet in a small group with one of our outstanding CAs (section leaders) and work through problems. If you have taken any of the CS106 classes, our sections will be very similar—except with more probability. Sign-ups for sections will go out on Thursday, January 9 and will be open until Saturday, January 11. We will let you know which section you are in by Monday, January 13, and you will have your first section that week (during Week 2). The full section attendance policy is on the course website (`http://cs109.stanford.edu/section/attendance.html`).

Grading

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

Assignments:	45%
Midterm Exam:	20%
Final Exam:	30%
Participation:	5%

Late Assignments

Since we anticipate that during the quarter, there may be unforeseen circumstances that make it difficult to turn in homework assignments on time, we will grant each student two free late days to be used when turning in assignments (problem sets). Each free late day counts as a **class** day. Therefore, a homework assignment due on Monday may be turned on Wednesday before 1:00pm 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. ***Note that no late days may be used beyond the last class day of the regular quarter (Friday March 13).***

After the two free late days are used, late assignments will be assessed a 20% reduction in score for each class day that they are turned in late. Moreover, no assignment will be accepted more than a week late (regardless of the use of free late days), since we will be releasing solutions to the assignments one week after their due date. Extensions beyond these free late days are rarely granted, and then, only for medical emergencies. Such extensions must be cleared prior to the assignment being due, and the request should be made directly to David.

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. The OAE is located at 563 Salvatierra Walk (`https://oae.stanford.edu`, phone: 723-1066).