ORIE 4520 - Stochastics at Scale

Instructor: Siddhartha Banerjee

Semester: Fall 2015

August 26, 2015

Essential Course Information

- *Instructor*
  Prof. Siddhartha Banerjee
  Office: 229 Rhodes Hall
  E-mail: `sbanerjee@cornell.edu`
  Website: people.orie.cornell.edu/sbanerjee/
  Office hours: MW 2:30pm-3:30pm (immediately after class)
- *Teaching Assistant*
  Anna Srapionyan
  E-mail: `as3348@cornell.edu`

Essential Course Information (contd.)

- *Lectures and Recitations*
  Course Number: ORIE 4520
  Class time: MWF 1:25-2:15pm
  Class location: Phillips 403
  Recitation time/location: To be decided
  (Recitation time on schedule: Tuesay, 2:55-4:10pm)
- *Course Communication*:
  Website: `http://people.orie.cornell.edu/sbanerjee/orie4520f15.html`
  I will use BlackBoard for all announcements (search for ORIE 4520)

Course Prerequisites:

- Basic probability (at the level of ORIE 3500): Random variables, conditional probability and expectation, common probability distributions and their properties (binomial, geometric, exponential, Poisson); simulations.

Course Prerequisites:

- Basic probability (at the level of ORIE 3500): Random variables, conditional probability and expectation, common probability distributions and their properties (binomial, geometric, exponential, Poisson); simulations.
- Stochastic processes, in particular, Markov chains (at the level of ORIE 3510). There will be a recitation session covering the essentials.

Course Prerequisites:

- Basic probability (at the level of ORIE 3500): Random variables, conditional probability and expectation, common probability distributions and their properties (binomial, geometric, exponential, Poisson); simulations.
- Stochastic processes, in particular, Markov chains (at the level of ORIE 3510). There will be a recitation session covering the essentials.
- Algorithms and graph theory: asymptotic (Big O) notation, basic algorithms (sorting, searching), LP

Course Prerequisites:

- Basic probability (at the level of ORIE 3500): Random variables, conditional probability and expectation, common probability distributions and their properties (binomial, geometric, exponential, Poisson); simulations.
- Stochastic processes, in particular, Markov chains (at the level of ORIE 3510). There will be a recitation session covering the essentials.
- Algorithms and graph theory: asymptotic (Big O) notation, basic algorithms (sorting, searching), LP
- Mathematical maturity

What is 'scaling'??

A warmup example: Balls in Bins

Courtesy: www.fixturescloseup.com

A warmup example: Balls in Bins

Suppose you throw m balls into n bins uniformly at random (u.a.r.)

- Assume n is very very large.
  Think of number of balls $m(n)$ as a function of n.

A warmup example: Balls in Bins

Suppose you throw m balls into n bins uniformly at random (u.a.r.)

- Assume n is very very large.
  Think of number of balls $m(n)$ as a function of n.

Three Questions

- How big should m be before every bin has at least one ball?
  (The 'Coupon-Collector Problem')

A warmup example: Balls in Bins

Suppose you throw m balls into n bins uniformly at random (u.a.r.)

- Assume n is very very large.
  Think of number of balls $m(n)$ as a function of n.

Three Questions

- How big should m be before every bin has at least one ball?
  (The 'Coupon-Collector Problem')
- How big should m be before some bin has at least two balls?
  (The 'Birthday Paradox')

A warmup example: Balls in Bins

Suppose you throw m balls into n bins uniformly at random (u.a.r.)

- Assume n is very very large.
  Think of number of balls $m(n)$ as a function of n.

Three Questions

- How big should m be before every bin has at least one ball? (The 'Coupon-Collector Problem')
- How big should m be before some bin has at least two balls? (The 'Birthday Paradox')
- If we choose $m = n$, how many balls are there in the most-loaded bin?

Balls in Bins: Scaling

We throw m balls into n bins uniformly at random (u.a.r.)

Three Questions

- How big should m be before every bin has at least one ball? (The 'Coupon-Collector Problem')
  Answer:
- How big should m be before some bin has at least two balls? (The 'Birthday Paradox')
  Answer:
- If we choose $m = n$, how many balls are there in the most-loaded bin?
  Answer:

Balls in Bins: Scaling

We throw m balls into n bins uniformly at random (u.a.r.)

Three Questions

- How big should m be before every bin has at least one ball?
  (The 'Coupon-Collector Problem')
  Answer: $\Theta(n \log n)$
- How big should m be before some bin has at least two balls?
  (The 'Birthday Paradox')
  Answer:
- If we choose $m = n$, how many balls are there in the most-loaded bin?
  Answer:

Balls in Bins: Scaling

We throw m balls into n bins uniformly at random (u.a.r.)

Three Questions

- How big should m be before every bin has at least one ball? (The 'Coupon-Collector Problem')
  Answer: $\Theta(n \log n)$
- How big should m be before some bin has at least two balls? (The 'Birthday Paradox')
  Answer: $\Theta(\sqrt{n})$
- If we choose $m = n$, how many balls are there in the most-loaded bin?
  Answer:

Balls in Bins: Scaling

We throw m balls into n bins uniformly at random (u.a.r.)

Three Questions

- How big should m be before every bin has at least one ball? (The 'Coupon-Collector Problem')
  Answer: $\Theta(n \log n)$
- How big should m be before some bin has at least two balls? (The 'Birthday Paradox')
  Answer: $\Theta(\sqrt{n})$
- If we choose $m = n$, how many balls are there in the most-loaded bin?
  Answer: $\Theta\left(\frac{\log n}{\log \log n}\right)$

Balls in Bins: Scaling

We throw m balls into n bins uniformly at random (u.a.r.)

Three Questions

- How big should m be before every bin has at least one ball? (The 'Coupon-Collector Problem')
  Answer: $\Theta(n \log n)$
- How big should m be before some bin has at least two balls? (The 'Birthday Paradox')
  Answer: $\Theta(\sqrt{n})$
- If we choose $m = n$, how many balls are there in the most-loaded bin?
  Answer: $\Theta\left(\frac{\log n}{\log \log n}\right)$

Takeaway: In large stochastic systems, simple questions have 'interesting' answers

Balls in Bins: One final twist

We throw m balls into n bins uniformly at random (u.a.r.)

- If we choose $m = n$, how many balls are there in the most-loaded bin?
  Answer: Maximum load is $\Theta\left(\frac{\log n}{\log\log n}\right)$

The power of two choices

Suppose instead we do the following:
For each ball, choose 2 bins u.a.r., and drop ball in less-loaded bin.

Balls in Bins: One final twist

We throw m balls into n bins uniformly at random (u.a.r.)

- If we choose $m = n$, how many balls are there in the most-loaded bin?
  Answer: Maximum load is $\Theta\left(\frac{\log n}{\log\log n}\right)$

The power of two choices

Suppose instead we do the following:
For each ball, choose 2 bins u.a.r., and drop ball in less-loaded bin.
The maximum load is now $\Theta(\log\log n)$

Balls in Bins: One final twist

We throw m balls into n bins uniformly at random (u.a.r.)

- If we choose $m = n$, how many balls are there in the most-loaded bin?
  Answer: Maximum load is $\Theta\left(\frac{\log n}{\log\log n}\right)$

The power of two choices

Suppose instead we do the following:
For each ball, choose 2 bins u.a.r., and drop ball in less-loaded bin.
The maximum load is now $\Theta(\log\log n)$

Takeaway: In large stochastic systems, small changes can lead to dramatic outcomes

A (tentative) list of topics

- **First unit:** Intro to randomized algorithms and scaling
  - Tools: Tail inequalities (the Chernoff bound), randomized rounding, random walks
  - Examples: Sorting, median finding, graph algorithms (min and max cut, centrality), routing problems

A (tentative) list of topics

- **First unit:** Intro to randomized algorithms and scaling
  - Tools: Tail inequalities (the Chernoff bound), randomized rounding, random walks
  - Examples: Sorting, median finding, graph algorithms (min and max cut, centrality), routing problems
- **Second unit:** Algorithms for dealing with 'big data'
  - Tools: Hashing, sketching, random projections
  - Examples: Basic operations for large data-sets, streaming data; algorithms for large graphs

A (tentative) list of topics

- **First unit:** Intro to randomized algorithms and scaling
  - Tools: Tail inequalities (the Chernoff bound), randomized rounding, random walks
  - Examples: Sorting, median finding, graph algorithms (min and max cut, centrality), routing problems
- **Second unit:** Algorithms for dealing with 'big data'
  - Tools: Hashing, sketching, random projections
  - Examples: Basic operations for large data-sets, streaming data; algorithms for large graphs
- **Third unit:** Threshold phenomena in large stochastic systems
  - Tools: Birth-death chains, branching processes, fluid approximations
  - Examples: Power of two choices, random graphs, epidemics

Back to Administrivia

Course Material

There is no required textbook for the course. I will cover different topics from different sources, and will periodically post notes and links to the relevant material on the website.

Back to Administrivia

Course Material

There is no required textbook for the course. I will cover different topics from different sources, and will periodically post notes and links to the relevant material on the website.

- Two good references for the first unit:
  - Randomized Algorithms by R. Motwani and P. Raghavan
  - Probability and Computing by M. Mitzenmacher and E. Upfal

Back to Administrivia

Course Material

There is no required textbook for the course. I will cover different topics from different sources, and will periodically post notes and links to the relevant material on the website.

- Two good references for the first unit:
  - Randomized Algorithms by R. Motwani and P. Raghavan
  - Probability and Computing by M. Mitzenmacher and E. Upfal
- Reference for the second unit:
  - Mining of Massive Datasets by J. Leskovec, A. Rajaraman and J. Ullman

Back to Administrivia

Course Material

There is no required textbook for the course. I will cover different topics from different sources, and will periodically post notes and links to the relevant material on the website.

- Two good references for the first unit:
  - Randomized Algorithms by R. Motwani and P. Raghavan
  - Probability and Computing by M. Mitzenmacher and E. Upfal
- Reference for the second unit:
  - Mining of Massive Datasets by J. Leskovec, A. Rajaraman and J. Ullman
- References for the third unit:
  - Networks, Crowds and Markets (Sections V, VI) by D. Easley and J. Kleinberg
  - Epidemics and Rumours in Complex Networks by M. Draief and L. Massoulié.

Coursework and Grading

- **Homework**:
  8 homeworks – weekly until the prelim, and biweekly after that. Homeworks due on Friday 12pm.

Coursework and Grading

- **Homework**:
  8 homeworks – weekly until the prelim, and biweekly after that. Homeworks due on Friday 12pm.
- **Exams**:
  One prelim: 90 min in-class exam, held during recitation hours (tentatively, during the week of 19th to 23rd October)
  No final exam (in place, we have a final project)

Coursework and Grading

- **Homework**:
  8 homeworks – weekly until the prelim, and biweekly after that. Homeworks due on Friday 12pm.
- **Exams**:
  One prelim: 90 min in-class exam, held during recitation hours (tentatively, during the week of 19th to 23rd October)
  No final exam (in place, we have a final project)
- **Project**:
  Read, simulate, research on chosen topic
  One-page proposal due Friday, October 23, 2015
  Deliverable: Paper (original research) or interactive document

Coursework and Grading

- **Homework**:
  8 homeworks – weekly until the prelim, and biweekly after that. Homeworks due on Friday 12pm.
- **Exams**:
  One prelim: 90 min in-class exam, held during recitation hours (tentatively, during the week of 19th to 23rd October)
  No final exam (in place, we have a final project)
- **Project**:
  Read, simulate, research on chosen topic
  One-page proposal due Friday, October 23, 2015
  Deliverable: Paper (original research) or interactive document
- **Grading**:
  Homeworks (45%) – $\max\{6 \times 5\% + 2 \times 10\%, 45\}$
  Prelim (25%), Project (25%+5%).