GB13604 - Maths for Computer Science

Lecture 10 – Probability, Part III

Claus Aranha
caranha@cs.tsukuba.ac.jp

College of Information Science

2018-12-19

Last updated December 16, 2018

This course is based on Mathematics for Computer Science, Spring 2015, by Albert Meyer and Adam Chlipala, Massachusetts Institute of Technology OpenCourseWare.

Introduction

- Law of Large Numbers and Sampling
- Random Walks and Probability Graphs
- Review Examination

Sampling

And the law of large numbers

Useful Statistics?

Consider a roll of a fair dice (from 1 to 6).

Expectation: The expected value of the dice is **3.5**. However, we will never roll **3.5**.

Probability: The probability of the value to be **6** is $\frac{1}{6}$. However, sometimes we can roll many times and never get a **6**.

So what are the meaning of these values?

The law of big numbers

For *n* rolls of a fair dice:

$$\Pr[\mathit{roll6}] = \frac{1}{6}$$

Means that, After many rolls, the fraction of 6 will be 1/6

For *n* rolls:

$$\frac{\text{\# 6 rolled}}{n} \rightarrow \frac{1}{6} \text{ as } n \rightarrow \infty$$

The law of big numbers

For n rolls:

$$\frac{\text{\# 6 rolled}}{n} \to \frac{1}{6} \text{ as } n \to \infty$$

Of course, for a non-infinite n, the proportion of 6 may be different than 1/6 in an unlucky roll.

However, for **big** n, this is unlikely. How unlikely?

Pr[Fraction of 6 = 1/6 $\pm x$ %]

# rolls	$\pm$10%	$\pm$5%
6	0.4	0.4
60	0.26	0.14
600	0.72	0.41
1200	0.88	0.56
3000	0.98	0.78
6000	0.999	0.98

- What is the probability that the fraction of 6 rolled is close to 1/6?
- Bigger with larger n
- Smaller with better precision.

- If you roll **3000** dice, and the # of 6 is not between 450 and 550, then you can be 98% confident that the dice us unfair
- If the # of 6 is not between 475 and 525, then you can be 98% confident that the dice us unfair

Fairness of dice

"If you roll **3000** dice, and the # of 6 is not between 450 and 550, then you can be 98% confident that the dice us unfair"

- The law of big numbers gives us bounds for the values of an independent random variable.
- It can be used to test the fairness (or correctness) of other random variables, such as Pseudo Random Number Generators!.
- This is **very** important for cryptographic applications.

Pairwise Independent Sampling

Theorem:
Let $R_1, R_2, R_3, \ldots, R_n$ be pairwise independent random variables, with the same finite mean μ and variance σ^2.

Let $A_n ::= R_1 + R_2 + \ldots + R_n/n$, then

$$\Pr[|A_n - \mu| > \delta] \leq \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2 \quad (1)$$

This theorem defines the probability of the mean of a **sample** (A_n) to be different from the mean of the **Random Variable** (μ) by a value δ.

(See the book for the derivation)

Sampling Experiments

Can you swim at the lake near the student plaza?

- According to EPA, a body of water is safe for swimming if its coliform count is $<$ 200, on average

  (coliform : 💩)

- How can you estimate the coliform average?

Sampling Experiments

Can you swim at the lake near the student plaza?

(estimating coliform count 💩)

- If you examine one place, you may get more than average, or less than average.
- Sampling Experiment: Make 32 experiments of coliform count at different locations in the lake.

Sampling Experiments: Results

- The average of the samples is 180
- But some of the samples are above 200!
- **Question:** Is the whole lake below 200?

Sampling Experiment and Parameter

Experiment Question: Is the estimated value based on 32 samples close to the real value?

- c ::= Real average coliform count in the lake.
- One Sample: Random variable with $\mu = c$
- n Samples: Mutually independent random variables with $\mu = c$
- A_n: average of the n samples (180)

Sampling Experiment and Parameter

Experiment Question: Is the estimated value based on 32 samples close to the real value?

Using the earlier **theorem**:

$$\Pr[|A_n - \mu| > \delta] \leq \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2$$

$$n = 32, \mu = c, \delta = 20$$

Sampling Experiment and Parameter

Experiment Question: Is the estimated value based on 32 samples close to the real value?

Using the earlier **theorem**:

$$\Pr[|A_n - \mu| > \delta] \leq \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2$$

$$n = 32, \mu = c, \delta = 20$$

But we don't know σ!!

Sampling Experiment and Parameter

Experiment Question: Is the estimated value based on 32 samples close to the real value?

Using the earlier **theorem**:

$$\Pr[|A_n - \mu| > \delta] \leq \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2$$

$$n = 32, \mu = c, \delta = 20$$

But we don't know σ!!
Let's assume a largest sample difference, $L = 50$

Experiment Probability Calculation

$$n = 32, \mu = c, \delta = 20, L = 50, \sigma = L/2 = 25$$

$$\Pr[|A_n - \mu| > \delta] \leq \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2$$

Experiment Probability Calculation

$$n = 32, \mu = c, \delta = 20, L = 50, \sigma = L/2 = 25$$

$$\Pr[|A_n - \mu| > \delta] \leq \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2$$

$$\Pr[|180 - c| > 20] \leq \frac{1}{32}\left(\frac{25}{20}\right)^2$$

Experiment Probability Calculation

$$n = 32, \mu = c, \delta = 20, L = 50, \sigma = L/2 = 25$$

$$\Pr[|A_n - \mu| > \delta] \leq \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2$$

$$\Pr[|180 - c| > 20] \leq \frac{1}{32}\left(\frac{25}{20}\right)^2$$

$$\Pr[|180 - c| > 20] \leq 0.05$$

Experiment Probability Calculation

$$n = 32, \mu = c, \delta = 20, L = 50, \sigma = L/2 = 25$$

$$\Pr[|A_n - \mu| > \delta] \leq \frac{1}{n}\left(\frac{\sigma}{\delta}\right)^2$$

$$\Pr[|180 - c| > 20] \leq \frac{1}{32}\left(\frac{25}{20}\right)^2$$

$$\Pr[|180 - c| > 20] \leq 0.05$$

$$\Pr[|180 - c| < 20] \geq 0.95$$

Confidence Interval

$$\Pr[|180 - c| < 20] \geq 0.95$$

We estimate with 95% confidence that the average coliform count is 180 ± 20.

Confidence Interval

$$\Pr[|180 - c| < 20] \geq 0.95$$

We estimate with 95% confidence that the average coliform count is 180 ± 20.

Be Careful!

- **Wrong interpretation**: There is a 0.95 probability that the average is between 160 and 200.
  (NO! The average is a fixed, real value!)

- **Correct interpretation**: Our sampling method estimates the average to be between 160 and 200. There is a 0.95 probability that our method is correct

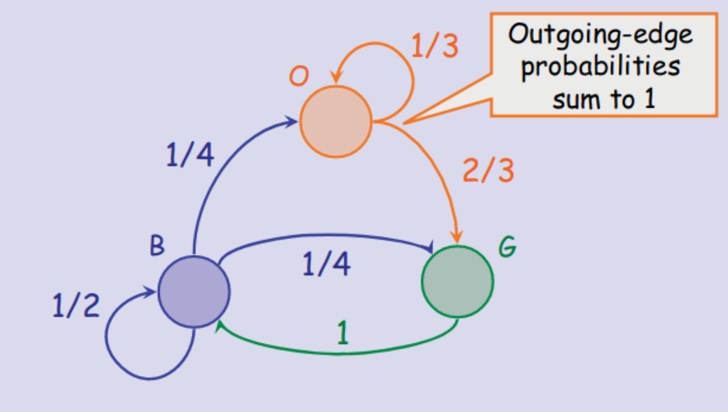
Random Walks

and

Probabilistic Graphs

Probabilistic Graph / State Machine

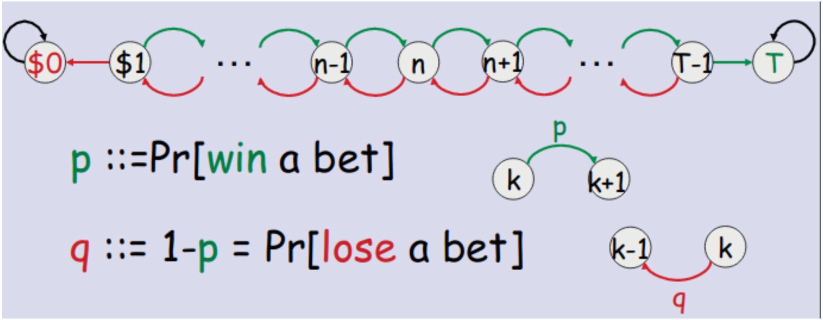
A probabilistic Graph (or State Machine), is a graph where each edge correspond to a transition probability.



Example: Gambler's Ruin

Imagine a game:

- With probability p you win 1\$.
- With probability $q = 1 - p$ you lose 1\$.
- You begin with n\$.
- What is the probability that you reach T\$ before you lose all?.



Applications of Random Walk

- **Physics:** Brownian Motion
- **Finance:** Stock prediction/simulation, options
- **Computer Science:** Web Search, Clustering

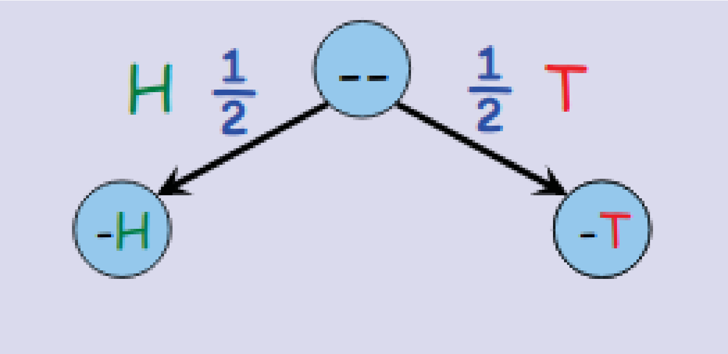
Example: Coin Game – HTH before TTH

Experiment: You throw a coin many times, and keep track of the last three results:

- If the sequence HTH happens before the sequence TTH, you win!
- If the sequence TTH happens before the sequence HTH, you lose!

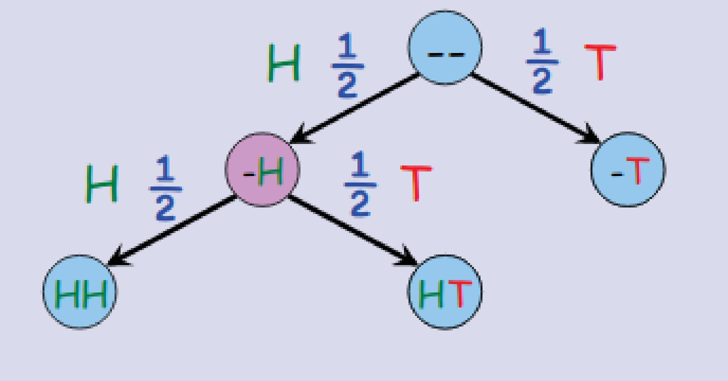
What is the probability that you win this game?

Example: Coin Game – HTH before TTH



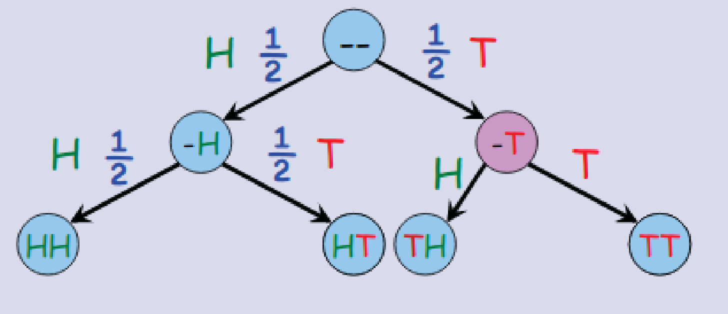
- Pr[Win] = Pr[Win|–] = 1/2 Pr[Win|-H] + 1/2 Pr[Win|-T]

Example: Coin Game – HTH before TTH



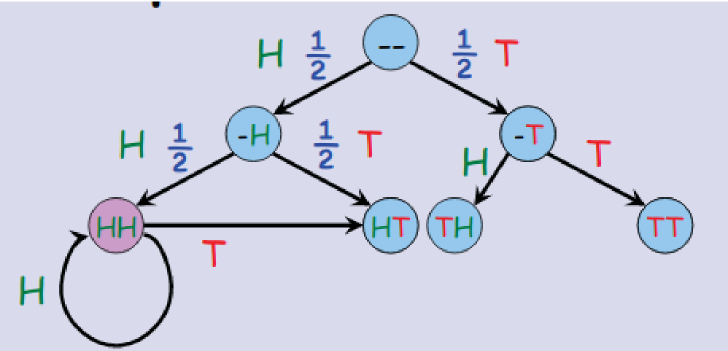
- Pr[Win|-H] = 1/2 Pr[Win|HH] + 1/2 Pr[Win|HT]

Example: Coin Game – HTH before TTH



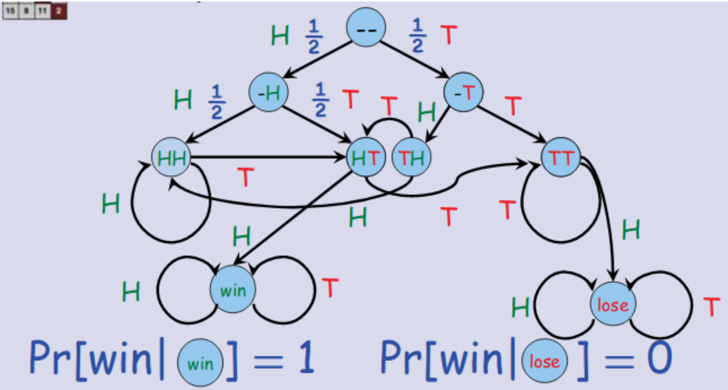
- Pr[Win|-T] = 1/2 Pr[Win|TH] + 1/2 Pr[Win|TT]

Example: Coin Game – HTH before TTH



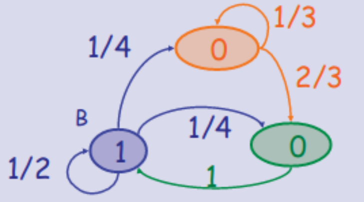
- Pr[Win|HH] = 1/2 Pr[Win|HH] + 1/2 Pr[Win|HT]

Example: Coin Game – HTH before TTH



And you can solve the system of linear equations for Pr[Win].

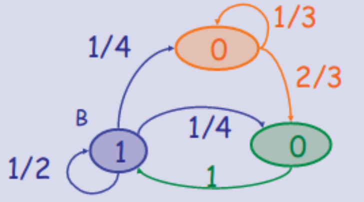
Stationary Distributions



Suppose you start at **B**: ($p_b = 1, p_o = 0, p_g = 0$)

What are the probabilities of each state: (p'_b, p'_o, p'_g) at the next step?

Stationary Distributions

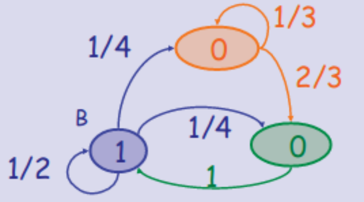


After 1 step, you follow the out-edges from B:

- $p'_b = p_b \cdot 1/2 = 1/2$
- $p'_o = p_b \cdot 1/4 = 1/4$
- $p'_g = p_g \cdot 1/4 = 1/4$

$(p'_b, p'_o, p'_g) = (1/2, 1/4, 1/4)$

Stationary Distributions



After 2 steps: (p''_b, p''_o, p''_g) from $(p'_b = 1/2, p'_o = 1/4, p'_g = 1/4)$

- $p''_b = p'_b \cdot 1/2 + p'_g \cdot 1 = 1/2$
- $p''_o = p'_b \cdot 1/4 + p'_o \cdot 1/3 = 5/24$
- $p''_g = p'_b \cdot 1/4 + p'_o \cdot 2/3 = 7/24$

$(p''_b, p''_o, p''_g) = (1/2, 5/24, 7/24)$

Edge Probability Matrix

The edge probability matrix is the same as the adjacency matrix, using edge probabilities instead of zeroes and ones.

$$M = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 0 & 1/3 & 2/3 \\ 1 & 0 & 0 \end{pmatrix}$$

We can use the edge probability matrix to calculate the walk state after i steps.

Edge Probability Matrix: Usage

You can use the edge probability matrix to calculate the state on the next step:

$$(p_b, p_o, p_g) \cdot M = (p'_b, p'_o, p'_g)$$

Stable Distribution

What is the graph state at step t?

$$(p_b, p_o, p_g) \cdot M^t = (p_b^t, p_o^t, p_g^t)$$

What is the graph state at step $t \to \infty$?

- Solve the system of equations:

$$\overrightarrow{s} \cdot M = \overrightarrow{s} \text{ and } \sum s_i = 1$$

Stable Distribution: Limitations

For some graphs, and some starting states, it may not be possible to find the stable distribution:

- Graph may not converge to a stable distribution
- Graph may have uncountable many stable distributions
- Graph may have multiple stable distributions

Google Webpage Ranking

- Which webpages are more important?
- Model of the Internet:
  Users click randomly on links in a webpage.
  Item sometimes the user starts over from a new page
- A page is “more important” if it is viewed more time.
  (probability in a “random walk”)

A random walk for the internet

- Represent the internet as a Directed Graph (DiGraph)
- Each webpage is a vertice, V_i
- A link from page V_i to page V_j is an Edge E_{ij}
- Identical probability for each edge out of V_i:
  $\Pr[E_{ij}] = 1/\deg(V)$

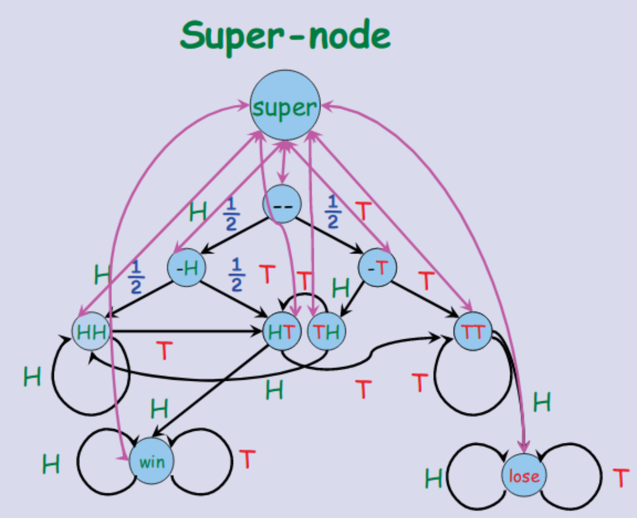
A random walk for the internet

To model starting over:

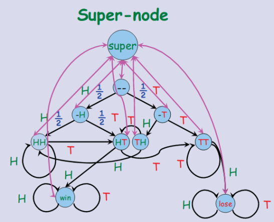
- Add a “super node” to the graph;
- The super node has an edge from it to every other node;
- Every other node has an edge back to the supernode;

(maybe with customized probabilities)

A random walk model of the internet



Pagerank



- Compute the stationary distribution $\vec{s}$

$$\text{Pagerank}(V) ::= s_V$$

- Rank page V_i above page V_j when:

$$s_V > s_W$$

Pagerank

Resistant to scamming:

- Creating fake nodes pointing to self does not help.
- Adding links to other nodes has Diminishing Returns

Importance of supernode:

- Ensures unique stable distribution $\vec{s}$
- Ensures that every initial condition $\vec{p}$ converges to $\vec{s}$

(Of course, Google's algorithm today has more tricks)

Exam Information

- Sample Exam
- Class Evaluation
- Final Exam
- Grades

Sample Exam & Class Evaluation

- The sample exam is an idea of what kind of questions are asked in the final exam.
- The sample exam **will not** be graded.
- Please feel free to ask for help in the sample exam.
- Please complete the Class Evaluation too.

Final Exam

- **You can bring:** 1 note page
  (A4, front and back, with name and student number)
- **You can bring:** Dictionary, Electronic Dictionary, calculator
- **You can NOT use:** Textbook, class slides, computer.

Grades

- Assignment Grade: Before the Final Exam
- Final Exam Grade: Before 1/4
- Grade Questions: Until 1/11