In-Class Problems Week 13, Fri.

The first three problems are carried over from Wednesday.

Problem 1.
A recent Gallup poll found that 35% of the adult population of the United States believes that the theory of evolution is "well-supported by the evidence." Gallup polled 1928 Americans selected uniformly and independently at random. Of these, 675 asserted belief in evolution, leading to Gallup's estimate that the fraction of Americans who believe in evolution is $675/1928 \approx 0.350$. Gallup claims a margin of error of 3 percentage points, that is, he claims to be confident that his estimate is within 0.03 of the actual percentage.

(a) What is the largest variance an indicator variable can have?

(b) Use the Pairwise Independent Sampling Theorem to determine a confidence level with which Gallup can make his claim.

(c) Gallup actually claims greater than 99% confidence in his estimate. How might he have arrived at this conclusion? (Just explain what quantity he could calculate; you do not need to carry out a calculation.)

(d) Accepting the accuracy of all of Gallup's polling data and calculations, can you conclude that there is a high probability that the number of adult Americans who believe in evolution is 35 ± 3 percent?

Problem 2.
Yesterday, the programmers at a local company wrote a large program. To estimate the fraction, b, of lines of code in this program that are buggy, the QA team will take a small sample of lines chosen randomly and independently (so it is possible, though unlikely, that the same line of code might be chosen more than once). For each line chosen, they can run tests that determine whether that line of code is buggy, after which they will use the fraction of buggy lines in their sample as their estimate of the fraction b.

The company statistician can use estimates of a binomial distribution to calculate a value, s, for a number of lines of code to sample which ensures that with 97% confidence, the fraction of buggy lines in the sample will be within 0.006 of the actual fraction, b, of buggy lines in the program.

Mathematically, the *program* is an actual outcome that already happened. The *sample* is a random variable defined by the process for randomly choosing s lines from the program. The justification for the statistician's confidence depends on some properties of the program and how the sample of s lines of code from the program are chosen. These properties are described in some of the statements below. Indicate which of these statements are true, and explain your answers.

1. The probability that the ninth line of code in the *program* is buggy is b.
2. The probability that the ninth line of code chosen for the *sample* is defective, is b.
3. All lines of code in the program are equally likely to be the third line chosen in the *sample*.
4. Given that the first line chosen for the *sample* is buggy, the probability that the second line chosen will also be buggy is greater than b.

5. Given that the last line in the *program* is buggy, the probability that the next-to-last line in the program will also be buggy is greater than b.

6. The expectation of the indicator variable for the last line in the *sample* being buggy is b.

7. Given that the first two lines of code selected in the *sample* are the same kind of statement—they might both be assignment statements, or both be conditional statements, or both loop statements,...—the probability that the first line is buggy may be greater than b.

8. There is zero probability that all the lines in the *sample* will be different.

Problem 3.
A defendent in traffic court is trying to beat a speeding ticket on the grounds that—since virtually everybody speeds on the turnpike—the police have unconstitutional discretion in giving tickets to anyone they choose. (By the way, we don't recommend this defense `:-)`.)

To support his argument, the defendent arranged to get a random sample of trips by 3,125 cars on the turnpike and found that 94% of them broke the speed limit at some point during their trip. He says that as a consequence of sampling theory (in particular, the Pairwise Independent Sampling Theorem), the court can be 95% confident that the actual percentage of all cars that were speeding is $94 \pm 4\%$.

The judge observes that the actual number of car trips on the turnpike was never considered in making this estimate. He is skeptical that, whether there were a thousand, a million, or 100,000,000 car trips on the turnpike, sampling only 3,125 is sufficient to be so confident.

Suppose you were were the defendent. How would you explain to the judge why the number of randomly selected cars that have to be checked for speeding *does not depend on the number of recorded trips*? Remember that judges are not trained to understand formulas, so you have to provide an intuitive, nonquantitative explanation.

Problem 4.
We want to store 2 billion records into a hash table that has 1 billion slots. Assuming the records are randomly and independently chosen with uniform probability of being assigned to each slot, two records are expected to be stored in each slot. Of course under a random assignment, some slots may be assigned more than two records.

(a) Show that the probability that a given slot gets assigned more than 23 records is less than e^{-36}.

Hint: For $c = 12$, the value of $c \ln c - c + 1$ is greater than 18.

(b) Show that the probability that there is a slot that gets assigned more than 23 records is less than e^{-15}. This is less than $1/3,000,000$. *Hint:* $\ln 10^9 < 21$.

The Chernoff Bound: Let T be the sum of a finite number of mutually independent variables whose codomain is the real interval $[0, 1]$. Then for all $c \geq 1$,

$$\Pr[T \geq c \operatorname{Ex}[T]] \leq e^{-\beta(c) \operatorname{Ex}[T]}$$

where $\beta(c) ::= c \ln c - c + 1$.

Problem 5.
In this problem you will check a proof of:

Theorem (Murphy's Law). *Let $A_1, A_2, \ldots A_n$ be mutually independent events, and let T be the number of these events that occur. The probability that none of the events occur is at most $e^{-\operatorname{Ex}[T]}$.*

To prove Murphy's Law, note that

$$T = T_1 + T_2 + \cdots + T_n, \tag{1}$$

where T_i is the indicator variable for the event A_i. Also, remember that

$$1 + x \leq e^x \tag{2}$$

for all x.

Justify each line in the following derivation (without looking it up in the text):

Proof.

$$\begin{aligned}
\Pr[T = 0] &= \overline{A_1 \cup A_2 \cup \cdots \cup A_n} \\
&= \Pr[\overline{A_1} \cap \overline{A_2} \cap \cdots \cap \overline{A_n}] \\
&= \prod_{i=1}^{n} \Pr[\overline{A_i}] \\
&= \prod_{i=1}^{n} 1 - \Pr[A_i] \\
&\leq \prod_{i=1}^{n} e^{-\Pr[A_i]} \\
&= e^{-\sum_{i=1}^{n} \Pr[A_i]} \\
&= e^{-\sum_{i=1}^{n} \operatorname{Ex}[T_i]} \\
&= e^{-\operatorname{Ex}[T]}.
\end{aligned}$$

■

Take 2

Redoing problems from ~~Mon~~ Wed
- well not enough time

1. Recent Gallup poll
   35% believe evolution
   1928 polled
   625 said ~~yes~~ believe
   3% margin of error
   So confident

a) what is largest var an indicator variable can have?

$E[\ ] = .5$

$\frac{1}{2}(1-.5)^2 + \frac{1}{2}(0-.5)^2 = \frac{1}{4}$

b) Pairwise Ind. Sampling Theorm to find confidence

Let $G_1, \ldots, G_n$ be pairwise ind μ = mean σ = dev

$S_n = \sum_{i=1}^{n} G_i$ ← just the sum

②

Then

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq x\right) \leq \frac{1}{n}\left(\frac{\sigma}{x}\right)^2$$

So $x = .03$

So bound $\frac{1}{1928}\left(\frac{1/2}{.03}\right)^2$ ↓ Use max σ $\frac{1}{2}$ - remember from 6.0(4(1))

$\approx .144 \approx 1/7$ ✓

c) How did he get 99%

He's lying
↑ what I thought

Could also say he used a smaller var than $\frac{1}{4}$

TA: But what is the justification?

d) Can you conclude there is a high prob # of americans is $35 \pm 3\%$

Well its that scemantic thing about wording

Prob our estimation procedure will give $35 \pm 3\%$

it either is or it isn't

(3)

2. Programmer at ~~FUBAR~~ company

b - fraction lines buggy

Chooses lines ind

- but lines are connected
  So this qu is bogus
  it could be typo in a particular line

Company can estimate binomial dist to calc a value s for # of line to sample

So 97% confidence is within .006 of actual fraction b

- (So this is Chebchev

$$P\left(\left|\frac{B_n}{B} - b\right| \geq .003\right) \leq .97$$

B: # of lines buggy

④ T/F?

1. Prob 9th line is buggy is b
 ✓ False - this is what I got wrong on
 tutor - its 0 or 1

2. Samples - ~~yes~~ No - its buggy or its not
 ~~[illegible]~~
 But sample chosen randomly, so True.

3. All lines = ly likly
 Yes
 ~~[illegible]~~ True

4. No ind

5. No - ind and pre chosen

6. Yes - E[ ] sample

7. Yes - since not ind - ~~[illegible]~~ ~~[illegible]~~ same ~~false~~ types

8. Are samples removed from program - Not 0 False
 ~~Board~~ Group agrees

5

(This is actually kind -of fun)

3. Traffic court

- Police have too much discretion

3125 cars

94% broke the law

So court can be 95% confident actual speeding is 94 ± 4%

Actual # trips not considered

? Doesn't matter

But how to say

Explain pairwise ind formula

Go back to Wed's lecture

? Because they are likely to be same

Picked cars at random before trip started

picked large enough to get idea

? WLLN? - gets close ← explain

In-Class 13 Fri Actual

5/6

4. 2 billion records in hash table w/ 1 bill slots
Records saved ind
2 records in each slot expected
but will actually be more or less

a) Show prob given slot gets assined > 23 records is < e^{-36}

Chebchev

$$P(X \geq 24) \leq e^{-(12 \ln 12 - 11)2}$$
$$\leq e^{-36}$$

b) Show P(slot assiged > 23 records) is less then e^{15}

Booles Inequality

$$P[E] \leq \sum_{i=1}^{10^9} P(X_i) = 10^9 e^{-36}$$
$$\leq e^{-15}$$

②

5. Check proof of Murphy's Law

(what does "Check" mean?

Justify each line

(I shall practice this)

Solutions to In-Class Problems Week 13, Fri.

The first three problems are carried over from Wednesday.

Problem 1.
A recent Gallup poll found that 35% of the adult population of the United States believes that the theory of evolution is "well-supported by the evidence." Gallup polled 1928 Americans selected uniformly and independently at random. Of these, 675 asserted belief in evolution, leading to Gallup's estimate that the fraction of Americans who believe in evolution is $675/1928 \approx 0.350$. Gallup claims a margin of error of 3 percentage points, that is, he claims to be confident that his estimate is within 0.03 of the actual percentage.

(a) What is the largest variance an indicator variable can have?

Solution.

$$\frac{1}{4}$$

By Lemma ??, $\text{Var}[H] = pq$.

Noting that $d\, p(1-p)/dp = 2p - 1$ is zero when $p = 1/2$, it follows that the maximum value of $p(1-p)$ must be at $p = 1/2$, so the maximum value of $\text{Var}[H]$ is $(1/2)(1-(1/2)) = 1/4$. ■

(b) Use the Pairwise Independent Sampling Theorem to determine a confidence level with which Gallup can make his claim.

Solution. By the Pairwise Independent Sampling, the probability that a sample of size $n = 1928$ is further than $x = 0.03$ of the actual fraction is at most

$$\left(\frac{\sigma}{x}\right)^2 \cdot \frac{1}{n} \leq \left(\frac{1}{4(0.03)^2} \cdot \frac{1}{1928}\right) \leq 0.144$$

so we can be confident of Gallup's estimate at the 85.6% level. ■

(c) Gallup actually claims greater than 99% confidence in his estimate. How might he have arrived at this conclusion? (Just explain what quantity he could calculate; you do not need to carry out a calculation.)

Solution. Gallup's sample has a binomial distribution $B_{1928,p}$ for an unknown p he estimates to be about 0.35. So he wants an upper bound on

$$\Pr[\left|\frac{B_{1928,p}}{1928} - p\right| > 0.03]$$

By part (a), the variance of $B_{n,p}$ is largest when $p = 1/2$, which suggests that the probability that a sample average differs from the actual mean will be largest when $p = 1/2$. This is in fact the case. So Gallup will calculate

$$\begin{aligned}\Pr[\left|\frac{B_{1928,1/2}}{1928} - \frac{1}{2}\right| > 0.03] &= \Pr[\left|B_{1928,1/2} - \frac{1928}{2}\right| > 0.03(1928)] \\ &= \Pr[906 \leq B_{1928,1/2} \leq 1021] \\ &= \frac{\sum_{i=906}^{1021} \binom{1928}{i}}{2^{1928}} \approx 0.9912.\end{aligned}$$

Mathematica will actually calculate this sum exactly. There are also simple ways to use Stirling's formula to get a good estimate of this value. ■

(d) Accepting the accuracy of all of Gallup's polling data and calculations, can you conclude that there is a high probability that the number of adult Americans who believe in evolution is 35 ± 3 percent?

Solution. No. As explained in Notes and lecture, the assertion that fraction p is in the range 0.35 ± 0.03 is an assertion of fact that is either true or false. The number p is a *constant*. We don't know its value, and we don't know if the asserted fact is true or false, but there is nothing probabilistic about the fact's truth or falsehood.

We *can* say that either the assertion is true or else a 1-in-100 event occurred during the poll. Specifically, the unlikely event is that Gallup's random sample was unrepresentative. This may convince you that p is "probably" in the range 0.35 ± 0.03, but this informal "probably" is not a mathematical probability. ■

Problem 2.

Yesterday, the programmers at a local company wrote a large program. To estimate the fraction, b, of lines of code in this program that are buggy, the QA team will take a small sample of lines chosen randomly and independently (so it is possible, though unlikely, that the same line of code might be chosen more than once). For each line chosen, they can run tests that determine whether that line of code is buggy, after which they will use the fraction of buggy lines in their sample as their estimate of the fraction b.

The company statistician can use estimates of a binomial distribution to calculate a value, s, for a number of lines of code to sample which ensures that with 97% confidence, the fraction of buggy lines in the sample will be within 0.006 of the actual fraction, b, of buggy lines in the program.

Mathematically, the *program* is an actual outcome that already happened. The *sample* is a random variable defined by the process for randomly choosing s lines from the program. The justification for the statistician's confidence depends on some properties of the program and how the sample of s lines of code from the program are chosen. These properties are described in some of the statements below. Indicate which of these statements are true, and explain your answers.

1. The probability that the ninth line of code in the *program* is buggy is b.

   Solution. False.

   The program has already been written, so there's nothing probabilistic about the buggyness of the ninth (or any other) line of the program: either it is or it isn't buggy, though we don't know which. You could argue that this means it is buggy with probability zero or one, but in any case, it certainly isn't b. ■

2. The probability that the ninth line of code chosen for the *sample* is defective, is b.

   Solution. True.

   The ninth line sampled is equally likely to be any line of the program, so the probability it is buggy is the same as the fraction, b, of buggy lines in the program. ■

3. All lines of code in the program are equally likely to be the third line chosen in the *sample*.

   Solution. True.

   The meaning of "random choices of lines from the program" is precisely that at each of the s choices in the sample, in particular at the third choice, each line in the program is equally likely to be chosen. ■

4. Given that the first line chosen for the *sample* is buggy, the probability that the second line chosen will also be buggy is greater than b.

   Solution. False.

   The meaning of "*independent* random choices of lines from the program" is precisely that at each of the s choices in the sample, in particular at the second choice, each line in the program is equally likely to be chosen, independent of what the first or any other choice happened to be. ■

5. Given that the last line in the *program* is buggy, the probability that the next-to-last line in the program will also be buggy is greater than b.

   Solution. False.

   As noted above, it's zero or one. ■

6. The expectation of the indicator variable for the last line in the *sample* being buggy is b.

   Solution. True.

   The expectation of the indicator variable is the same as the probability that it is 1, namely, it is the probability that the sth line chosen is buggy, which is b, by the reasoning above. ■

7. Given that the first two lines of code selected in the *sample* are the same kind of statement—they might both be assignment statements, or both be conditional statements, or both loop statements,...—the probability that the first line is buggy may be greater than b.

   Solution. True.

   We don't know how prone to bugginess different kinds of statements may be. It could be for example, that conditionals are more prone to bugginess than other kinds of statements, and that there are more conditional lines than any other kind of line in the program. Then given that two randomly chosen lines in the sample are the same kind, they are more likely to be conditionals, which makes them more prone to bugginess. That is, the conditional probability that they will be buggy would be greater than b. ■

8. There is zero probability that all the lines in the *sample* will be different.

   Solution. False.

   We know the length, r, of the program is larger than the "small" sample size, s, in which case the probability that all the lines in the sample are different is

$$\frac{r}{r} \cdot \frac{r-1}{r} \cdot \frac{r-2}{r} \cdots \frac{r-(s-1)}{r} = \frac{r!}{(r-s)!\,r^s} > 0.$$

   Of course it would be true by the Pigeonhole Principle if $s > r$. ■

Problem 3.

A defendent in traffic court is trying to beat a speeding ticket on the grounds that—since virtually everybody speeds on the turnpike—the police have unconstitutional discretion in giving tickets to anyone they choose. (By the way, we don't recommend this defense :-).)

To support his argument, the defendent arranged to get a random sample of trips by 3,125 cars on the turnpike and found that 94% of them broke the speed limit at some point during their trip. He says that as a consequence of sampling theory (in particular, the Pairwise Independent Sampling Theorem), the court can be 95% confident that the actual percentage of all cars that were speeding is $94 \pm 4\%$.

The judge observes that the actual number of car trips on the turnpike was never considered in making this estimate. He is skeptical that, whether there were a thousand, a million, or 100,000,000 car trips on the turnpike, sampling only 3,125 is sufficient to be so confident.

Suppose you were were the defendent. How would you explain to the judge why the number of randomly selected cars that have to be checked for speeding *does not depend on the number of recorded trips*? Remember that judges are not trained to understand formulas, so you have to provide an intuitive, nonquantitative explanation.

Solution. This was intended to be a thought-provoking, conceptual question. In past terms, although most of the class could follow the derivations and crank through the formulas to calculate sample size and confidence levels, many students couldn't articulate, and indeed didn't really believe that the derived sample sizes were actually adequate to produce reliable estimates.

Here's a way to explain why we model sampling cars as independent coin tosses that might work, though we aren't sure about this.

> Of the approximately 36,000,000 recorded turnpike trips by cars in 2009, there were some *unknown* number, say 35,000,000, that broke the speed limit at some point during their trip. So in this case, the *fraction* of speeders is 35,000,000/36,000,000 which is a little over 0.97.
>
> To estimate this unknown fraction, we randomly select some trip from the 36,000,000 recorded in such a way that *every trip has an equal chance of being picked.* Picking a trip to check for speeding this way amounts to rolling a pair dice and checking that double sixes were not rolled—this has exactly the same probability as picking a speeding car.
>
> After we have picked a car trip and checked if it ever broke the speed limit, make another pick, again making sure that every recorded trip is equally likely to be picked the second time, and so on, for picking a bunch of trips. Now each pick is like rolling the dice and checking against double sixes.
>
> Now everyone understands that if we keep rolling dice looking for double sixes, then the longer we roll, the closer the fraction of rolls that are double sixes will be to 1/36, since only 1 out of the 36 possible dice ouotcomes is double six. Mathematical theory lets us calculate us how many times to roll the dice to make the fraction of double sixes very likely close to 1/36, but we needn't go into the details of the calculation.
>
> Now suppose we had a different number of recorded trips, but the same fraction were speeding. Then we could simply use the same dice in the same way to estimate the speeding fraction from this different set of trip records.
>
> So the number of rolls needed does not depend on how many trips were recorded, it just depends on the fraction of recorded speeders.

■

Problem 4.
We want to store 2 billion records into a hash table that has 1 billion slots. Assuming the records are randomly and independently chosen with uniform probability of being assigned to each slot, two records are expected to be stored in each slot. Of course under a random assignment, some slots may be assigned more than two records.

(a) Show that the probability that a given slot gets assigned more than 23 records is less than e^{-36}.
Hint: For $c = 12$, the value of $c \ln c - c + 1$ is greater than 18.

Solution. Let T be the number of records assigned to a particular slot, say the first one. So $\text{Ex}[T] = 2$. Then by Chernoff

$$\Pr[T \geq 24] = \Pr[T \geq 12\,\text{Ex}[T]] \leq e^{-\beta(12)\,\text{Ex}[T]} < e^{-18 \cdot 2} = e^{-36}.$$

■

(b) Show that the probability that there is a slot that gets assigned more than 23 records is less than e^{-15}. This is less than $1/3,000,000$. *Hint:* $\ln 10^9 < 21$.

Solution. By the Union Bound, the probability that some slot gets assigned more than 23 records is at most 1 billion times the probability that each particular slot gets assigned more than 23 records, and is therefore

$$\leq 10^9 \cdot e^{-36} < e^{21} \cdot e^{-36} = e^{-15} < \frac{1}{3,270,000} < \frac{1}{3,000,000}.$$

■

The Chernoff Bound: Let T be the sum of a finite number of mutually independent variables whose codomain is the real interval $[0, 1]$. Then for all $c \geq 1$,

$$\Pr[T \geq c\,\text{Ex}[T]] \leq e^{-\beta(c)\,\text{Ex}[T]}$$

where $\beta(c) ::= c \ln c - c + 1$.

Problem 5.
In this problem you will check a proof of:

Theorem (Murphy's Law). *Let $A_1, A_2, \ldots A_n$ be mutually independent events, and let T be the number of these events that occur. The probability that none of the events occur is at most $e^{-\text{Ex}[T]}$.*

To prove Murphy's Law, note that

$$T = T_1 + T_2 + \cdots + T_n, \tag{1}$$

where T_i is the indicator variable for the event A_i. Also, remember that

$$1 + x \leq e^x \tag{2}$$

for all x.

Justify each line in the following derivation (without looking it up in the text):

Solution. *Proof.*

$$
\begin{aligned}
\Pr[T=0] &= \overline{A_1 \cup A_2 \cup \cdots \cup A_n} && \text{(def. of } T\text{)}\\
&= \Pr[\overline{A_1} \cap \overline{A_2} \cap \cdots \cap \overline{A_n}] && \text{(De Morgan's law)}\\
&= \prod_{i=1}^{n} \Pr[\overline{A_i}] && \text{(mutual independence of } A_i\text{'s)}\\
&= \prod_{i=1}^{n} 1 - \Pr[A_i] && \text{(complement rule)}\\
&\le \prod_{i=1}^{n} e^{-\Pr[A_i]} && \text{(by (2))}\\
&= e^{-\sum_{i=1}^{n} \Pr[A_i]} && \text{(exponent algebra)}\\
&= e^{-\sum_{i=1}^{n} \mathrm{Ex}[T_i]} && \text{(expectation of indicator variable)}\\
&= e^{-\mathrm{Ex}[T]}. && \text{((1) \& linearity of expectation)}
\end{aligned}
$$

■

TP12 Deviation

TP.12.1 Suppose 5 is expected

$X = RV > 0$

$E[X] = 5$

a) which statement is true?
At least one value

$= 0$

≤ 2.5

< 5

≤ 5

> 10

~~less than = at most~~

oh

$<$ less than $\leq$ at most

what if 5, 5, 5?

≤ 5 ✓

b) $E[x^2]$ is

$= 25$

$\neq 25$ ← don't know - could be if 5, 5, 5

≤ 25

≥ 100 ← could be - but non reg ? so can it ✓

$\leq E[X]$

② TP 12.2 Above avg # of Fingers

90% drivers think they are above avg

the ~~fus~~ vast majority of people hve ≥ avg # fingers

1 X

2 X

3 ✓

4. ✓ just restates claim

5 X

6 ✓) does not matter how much larger

7 X ? tricky!

3 4 6 ~~X~~ X

3 4 5 6 X

~~3M~~ 4 5 7 X

3 4 7 X

I hate these type of qu!

367 or 67

3

JP 123 Expectation of x^2

X = RV uniform $[-n, n]$

$Y = x^2$

What is true

1 ✓ uniform

2 ✓ No since all values non neg ← the big fact I forgot

3 ⊗

4 ✓ linearity

5 ✓ I think, since 0

6 x

7 x

1245 ⊗

145 ⊗

14 ⊗

124 ⊗

134 ⊗

1345

(4)

TP 12.4 Practice w/ Bounds

120 students take final

mean $= 90$

Can't assume final out of 100

a) State best possible upper bound on

\# students $\rightarrow \geq 180$ grade

Guessing Markov since 1st -1

But why can't it be Chebchev?

$$P(G \geq 180) \leq \frac{E[G]}{x}$$

\# of students

Grade

$$\frac{90}{180} = \frac{1}{2}$$

Oh \# students $\frac{1}{2} \cdot 120 = 60$

read carefully!

b) Now lowest score $= 30$

$T = G - 30$

$$P(T \geq 150) \leq \frac{E[G] - 30}{x} = \frac{60}{150}$$

$\frac{6}{15} \cdot 120 = 48$

5

TP. 12.5 Flipping Coins

Suppose flip fair coin 100 times - mutually ind

1. $E[\#\text{ heads}] = 50$

2. Upper bound # heads ≥ 70

$$P(H \geq 70) \leq \frac{E[\text{heads}]}{70}$$

$$= \frac{50}{70} = 5/7$$

3. Var on # heads

? what is best way

$$\text{var}(H) = E[H^2] - E[H]^2$$

is this easy to find

$$\underbrace{\left(\tfrac{1}{2}\right)^{100}}_{\text{same}} \cdot 100^2 + \left(\tfrac{1}{2}\right)^{99} \tfrac{1}{2} \cdot 99^2$$

$$\sum_{x=0}^{100} x^2$$

WA $= 338350$

No $\sum_{x=0}^{100} \left(\tfrac{1}{2}\right)^{100} \cdot x^2$

even worse

(6) or Add variances
— since ind

Var X_i = ?? how to find var manually ?

$\Sigma X - \mu$

$1 - .5 + 0 - .5$

$.5 - .5$

No

~~EV~~

$Var = E[(R - E[R])^2]$

$(1-.5)^2 + (0-.5)^2$

$\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

but E so

$\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{4}$

$\frac{1}{4}$ ✓

which is var of Bernoulli $p(1-p) = \frac{1}{4}$ here

So now add

$Var(H) = 100 \cdot \frac{1}{4} = 25$ ✓

(Need to learn all the ways to write stuff!)

7

4. What is upper bound # heads < 30
> 70

Chebchev

$\cancel{P(|H - 20|}$
$\uparrow$
~~more than 20 from mean~~

$$P(|H - 50| \geq 20) \leq \frac{Var[R]}{(c\sigma_R)^2}$$

$\uparrow$
$c\sigma_R$
(which doesn't matter here)

$$= \frac{25}{20^2}$$

$$= 1/16$$ ✓

Now I get it a little better!

8

TP 12.6 Random Sampling

Work for prez

Want fraction of voters p that ~~want to~~ vote for him

Select n voters ind

P = fraction they say

Part 1 Facts

Can calc how confident we are that

RV P takes a value near constant p

Which facts are true?

1. ✓ - Yes a random voter } The choice of a given voter is
2. X Not the prob } 1 or 0
3. When ever picked > 1, Guessing ✓
4. Can ~~b~~ voters be picked > 1? ✓
5. X No ind
6. ✓ Yes since same state

1 3 4 6 X   1 4 6 X

1 6 X   2 4 6

1 3 6 X

(9)

Part 2 What do you say?

$$P(|P-p| \le .04) \ge .95$$

$P = .53$

What do you say

1 x

2 x

3 ✓

4 ✓

3 4 ✓

(20)

T.P.12.7 Spiders + Flies

Spider is expecting guests

Wants 500 flies for dinner

100 flys pass by each hr

60 caught w/ $P() = \frac{1}{4}$

40 $\frac{3}{4}$

) ind

Spider only has $\frac{1}{100{,}000}$ chance to catch in 10 hrs

Part 1 Methods

- binomial ?
  - 1 if caught
  - look at density for 10 hrs ✓
- Markov ✓
- Chebyshev P(← no deation from mean ✗
- Chernoff – don't remember reading
  - – was the lotto - pick low prob #

Oh that section was ~~never~~ not published! when I printed

(1)

Can solve many problems

Certain RVs unlikely to exceed expectations

Exponential bound

$$P[T \geq cE[x]]] \leq e^{-kE[T]}$$

So can use?

No check buttons

Binomial ~~x~~ No - sum of binomial dist w/ diff p not binomial

Sum of 1000 Bernalli

Markov No - absorb over estimate

Chebyshev Yes

Chernoff Yes

Part 2 Which best

Chernoff or binomial?

they said this was very good

should be actual so best

No its Chernoff

(12)

c) To verify write formula ~~size~~ for spider's chances

Why $e^\wedge$ when have we done this before?

Just guess ~~Oh~~

Oh its Chernoff

$$e^{-k\, E[T]}$$ ← each chance

$k = c\ln(c) - c + 1$

↑ # of flies

So I think I get it now

So $E[T] = 1$ fly is caught

$$\frac{60}{100}\cdot\frac{1}{6}\cdot 1 + \frac{40}{100}\cdot\frac{3}{4}\cdot 1$$

$= .4$

~~And x is~~ oh T is for all

So $10\text{ hrs}\cdot 100\cdot .4 = 400$

(3)

Now c is multiplier

$P(T \geq c\,E[T])$

$cE[T] = 500$

$c\,400 = 500$

$c = \frac{5}{4}$

$k = \frac{5}{4} \ln\left(\frac{5}{4}\right) - \frac{5}{4} + 1$

So $e^{-\left(\frac{5}{4}\ln\left(\frac{5}{4}\right) - \frac{5}{4} + 1\right)400}$

is $P[T \geq cE[T]]$

$P[T \geq 500]$

5/9

Mathematics for Computer Science
MIT 6.042J/18.062J

Very Great Expectations, Gambler's Ruin

Repeating Tails

Flip a fair coin until a head;
F ::= #tails. If flip TTTH then F =3.
Flip again until head. If flip fewer than F tails, repeat.

Repeating Tails

1st try: TTH, must repeat
2nd try: H, must repeat
3rd try: TH, must repeat
4th try: TTTH, done!
R ::= #repeats R = 4 here

Repeating Tails

E[R|F=k] =
mean time to flip $T^k = \frac{1}{\Pr[T^k]} = 2^k$

$$E[R] = \sum_k E[R \mid F=k] \cdot \Pr[F=k]$$

Repeating Tails

E[R|F=k] =
mean time to flip $T^k = \frac{1}{\Pr[T^k]} = 2^k$

$$E[R] = \sum_k \frac{1}{\Pr[T^k]} \cdot \underbrace{\Pr[F=k]}_{2^{-(k+1)}}$$

Repeating Tails

E[R|F=k] =
mean time to flip $T^k = \frac{1}{\Pr[T^k]} = 2^k$

$$E[R] = \sum_k 2^k \cdot 2^{-(k+1)} = \sum_k \frac{1}{2} = \infty$$

Infinite Expectation

Can't use Law of Large Nums
what does sample data look like?
Infrequent large nums increase the average.
But if E[R] = ∞, maybe $E\left[\sqrt{R}\right] < \infty$

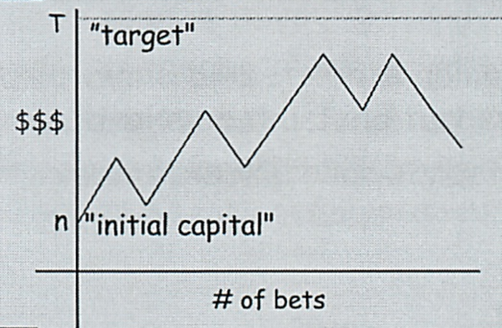
Infinite Expectation

Problems 1--3

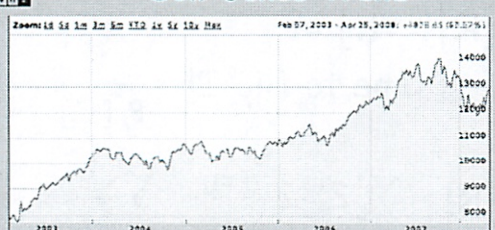
Gambler's Ruin

- Place \$1 bets until going broke or reaching target
- What is Pr[reach target]?

Gambler's Ruin



Dow Jones Trend



random steps with "up" bias?

Gambling: Fair Case

Suppose we're playing a fair game:
- Pr[win bet] = 1/2.

What is Pr[reach \$200] if we start with \$100? $\boxed{1/2}$

What about Pr[reach \$600] if we start with \$500? $\boxed{5/6}$

Gambling: Fair Case

In general, if we start with \$n

Pr[reach \$T] = n/T

What about an unfair game?

Gambling: Slightly Unfair

Betting **black** in US roulette

Pr{win bet} = 18/38 = 9/19 < 1/2

US Roulette

What is Pr[reach \$500+100] starting with \$500? (5/6 when fair)

< 1 / 37,000

What is Pr[reach \$1,000,100] starting with \$1,000,000? (≈ 1 when fair)

< 1 / 37,000

no matter how many \$ at start!

Gambler's Ruin

Parameters

- p ::= Pr{win \$1 bet}
- n ::= initial capital
- T ::= gambler's target

What is Pr[reach target]?

Gambler's Ruin

In fair game ($p = \frac{1}{2}$):

$$\Pr[\text{win}] = \frac{n}{T}$$

Gambler's Ruin

In unfair game ($p < \frac{1}{2}$):

$$\Pr[\text{win}] < \left(\frac{p}{q}\right)^{T-n}$$

intended profit

Profit $100 in US Roulette

$$\frac{p = \frac{18}{38}}{q = \frac{20}{38}} = \frac{9}{10}$$

$\Pr[\text{Profit } \$100] < (9/10)^{100}$

$< 1/37{,}648$

Team Problems

Problem 4

(5 min late)

∞ Expectation

(missed game description)

$E[R \mid F = k]$ = mean time to flip T^k

$$= \frac{1}{P[T^k]} = 2^k$$

$$E[R] = \sum_k \underbrace{\frac{1}{P[T^k]}}_{2^k} \cdot \underbrace{P[F=k]}_{2^{-(k+1)}}$$

this is this

$$= \sum_k 2^k \cdot 2^{-(k+1)} = \sum_k \frac{1}{2} = \infty$$

very divergent

Can't use law of large #

- avg of many trials
- since is no expectation
- since infreq large # ? the avg

(2)

But if $E[R] = \infty$, maybe $E[\sqrt{R}] < \infty$

Inclass 14 Mon

5/9

1. Biased coin w/ nonzero $p < 1$ heads
Toss until heads. then tossing till get long run of tails. Long run ~~means~~ = run of tails that is within 10 of initial run.

Prove $E[\#$ times toss head + start over$] = \infty$

: isn't this ~~first problem~~:
the lecture problem:

Can be within 10 - as opposed to being the exact #

Just replace 2^k w/ 2^{k-10}

2. Prove R is RV such that

$$PDF_R(n) = \frac{1}{Cn^3}$$

$$C = \sum_{n=1}^{\infty} \frac{1}{n^3}$$

a) Prove $E[R]$ is finite

$$E[R] = \sum_{n=\cancel{\infty}\, N}^{\infty} n \cdot PDF_R(n) = \sum_{n=1}^{\infty} n \cdot \frac{1}{Cn^3} = \frac{1}{C} \sum_{n=1}^{\infty} \frac{1}{n^2}$$

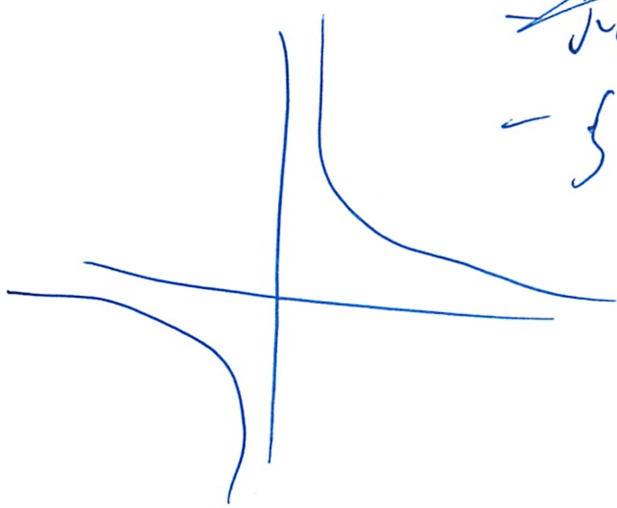
which is finite since $\int \frac{1}{n^2}$ is $\frac{1}{n}$ which goes to 0 ;;

③

b) Prove $Var(R)$ is ~~\~~ infinite

$$Var(R) = \sum_{n=1}^{\infty} n^2 PDF_R(n) - E[R]^2 = \frac{1}{C}\sum_{n=1}^{\infty}\frac{1}{n} - E[R]$$

which diverges



~~Just barly~~

$\sum \frac{1}{n}$ is $\ln n$

which diverges just bairly

(9)

3. Let T be (+) int RV so

$$PDF_T(n) = \frac{1}{an^2}$$

$$a = \sum_{a \in \mathbb{Z}^+} \frac{1}{n^2}$$

a) Prove $E[T]$ is infinite

$$E[T] = \sum_{n=1}^{\infty} n \, PDF_T(n) = \sum_{n=1}^{\infty} n \frac{1}{an^2} = \sum_{n=1}^{\infty} \frac{1}{an}$$

which diverges

b) Prove $E[\sqrt{T}]$ is infinite

$$E[\sqrt{T}] = \sum_{n=1}^{\infty} \sqrt{n} \, PDF_T(n) = \sum_{n=1}^{\infty} \sqrt{n} \frac{1}{an^2}$$

$$= \sum_{n=1}^{\infty} \frac{1}{an^{3/2}} \quad \text{converges}$$

$\frac{1}{n^{\text{any} > 1}}$ converges

try integrate it + look for if it goes to ∞

(5)

1 or board)

$P(\text{initial coin has } k \text{ tails}) = \left(\frac{1}{2}\right)^{k+1}$

$P(\text{second coin } J \text{ tails}) = \left(\frac{1}{2}\right)^{J+1}$

Second coin has $k-10 \leq J \leq J+10$ tails

$$P(\ J\ ) = \sum_{j=k-10}^{k+10} \left(\frac{1}{2}\right)^{J+1}$$

$$\left(\frac{1}{2}\right)^{k-10} \leq P(\text{second coin}) \leq \left(\frac{1}{2}\right)^{k-9}$$

$E[W] = P(\text{win w/ } k)\, k$

$$= \left(\frac{1}{2}\right)^{k+1} \sum_{j=k-10}^{k+10} \left(\frac{1}{2}\right)^{J+1} (J+k) > \left(\frac{1}{2}\right)^{k+1} \sum_{J=k-10}^{k+10} \left(\frac{1}{2}\right)^{J+1}$$

$$= \left(\frac{1}{2}\right)^{k+1} \sum_{k=k-10}^{J+10} \left(\frac{1}{2}\right)^{J+1}$$ since $(J+k) > 1$

$$= \left(\frac{1}{2}\right)^{k+1} \cdot \frac{1-\left(\frac{1}{2}\right)^{20}}{1-\frac{1}{2}}\, k$$

TAs didn't really understand qu much

In-Class Problems Week 14, Mon.

Problem 1.
You have a biased coin with nonzero probability $p < 1$ of coming up heads. You toss until a head comes up, and then, as in Section 18.8, you keep tossing until you get a long run of tails, but this time let "long run" mean a run of tails that is with 10 of the length your initial run. Prove that the expected number of times you toss a head and start over is still infinite.

within

Problem 2.
Let R be a positive integer valued random variable such that

$$\mathrm{PDF}_R(n) = \frac{1}{cn^3},$$

where

$$c ::= \sum_{n=1}^{\infty} \frac{1}{n^3}.$$

(a) Prove that $\mathrm{Ex}[R]$ is finite.

(b) Prove that $\mathrm{Var}[R]$ is infinite.

Problem 3.
Let T be a positive integer valued random variable such that

$$\mathrm{PDF}_T(n) = \frac{1}{an^2},$$

where

$$a ::= \sum_{n \in \mathbb{Z}^+} \frac{1}{n^2}.$$

(a) Prove that $\mathrm{Ex}[T]$ is infinite.

(b) Prove that $\mathrm{Ex}[\sqrt{T}]$ is finite.

Problem 4.
In gambler's ruin scenario, the gambler makes independent \$1 bets, where the probability of winning a bet p and of losing is $q ::= 1 - p$. The gambler keeps betting until he goes broke or reaches a target of T dollars.

Suppose $T = \infty$, that is, the gambler keeps playing until he goes broke. Let r be the probability that starting with $n > 0$ dollars, the gambler's stake ever gets reduced to $n - 1$ dollars.

(a) Explain why

$$r = q + pr^2.$$

(b) Conclude that if $p \leq 1/2$, then $r = 1$.

(c) Conclude that even in a fair game, the gambler is sure to get ruined *no matter how much money he starts with*!

Hint: If r_n is probability of ruin starting with stake n, then $r_n = r_{n+1}p + r_{n-1}q$, so

$$r_{n+1} = \frac{r_n}{p} - r_{n-1}\frac{q}{p}. \tag{1}$$

(d) Let t be the expected time for the gambler's stake to go down by 1 dollar. Verify that

$$t = q + p(1 + 2t).$$

Conclude that starting with a 1 dollar stake in a fair game, the gambler can expect to play forever!

(4)

Perfect final exam qv

1 again). $E[\text{cons total}] = \sum_{n=0}^{\infty} E[\text{cons } k]\, p(k)$

$$= \sum_{n=0}^{\infty} q^{10-k}\, pq^{k}$$

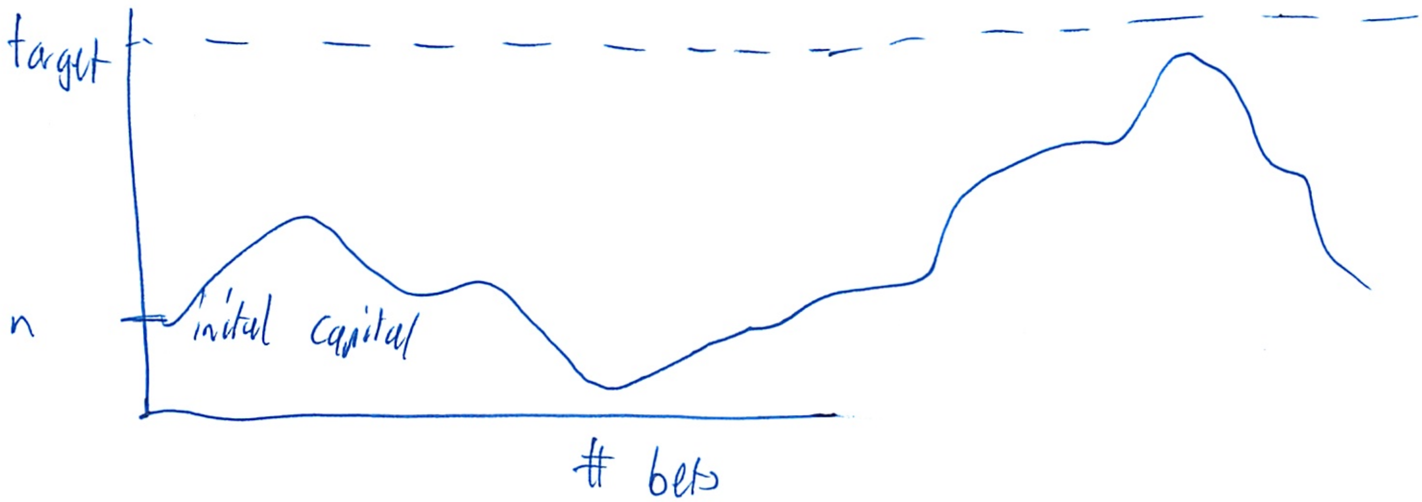
$$= \sum_{n=0}^{\infty} pq^{10} \to \infty$$

Gambler's Ruin

Place \$1 bets till broke or reach target

What is P(reach target)

- and not go bankrupt



Profit = T - N

Bankrupt

T-N

-N

continue oscilating P( ) = 0

- used to be class problem

(2)

Could also look at Dow
- are there any patterns?
- Quant Trading firms try

$P[\text{win bet}] = \frac{1}{2}$

$P[\text{reach } 200]$ if start 100
- You are right b/w the 2 boundries
- So by symmetry its $\frac{1}{2}$

$P[\text{reach } \cancel{800}\ 600]$ start 500

$= \frac{5}{6}$

$P[\text{reach } T]$ start n

$= \frac{n}{T}$ w/ fair game

So more lilty to hit goal w/ larger stake

---

but unfair game is totally different

Roulette
- house wins on green

$P(\text{win}) = \frac{18}{38}$

③

P(reach ~~\$~~ 500 +100) starting w/ 500

$\approx < \frac{1}{37{,}000}$

remember was $\frac{1}{6}$ when fair

P(~~win~~ reach ~~\$~~ 1,000,100) start w/ 1,000,000

was almost 1 when fair!

but still $\approx < \frac{1}{37{,}000}$

$p = p(\text{win } \$1 \text{ bet})$

n = inital capital

T = target

fair

$p = \frac{n}{T}$

unfair

$< \left(\frac{T}{q}\right)^{T-n}$

So Roullet

$$\left(\frac{\frac{18}{38}}{\frac{20}{38}}\right)^{100} = \left(\frac{9}{10}\right)^{100} \approx \frac{1}{37{,}648}$$

(4)

Used to derive formulas in prior classes
Nice proof in notes

(4A)

4. \$1 bets

p = winning

$1-q$

r = prob(starting w/ $n>0$ \$, the gamblers stake is ever reduced to $n-1$)

a) Why $r = \underbrace{q}_{\text{Prob lose 1 imediate}} + \underbrace{pr^2}_{\text{prob gain one then eventually lose two}}$

b) Conclude that if $p \le \frac{1}{2}$ then $r=1$

Just plug in a #

c) Conclude gambler is sure to be ruined, no matter how much \$ they start w/

r_n is prob of ruin, starting w/ n, then $r_n = r_{n+1}p + r_{n-1}q$

So $r_{n+1} = \frac{r_n}{p} - r_{n-1}\frac{q}{p}$

②

Recursive

Use $p = q = \frac{1}{2}$

$r_0 = 1$.

↑ already ruined

$r_1 = 1$

$r_2 = 2r_1 - r_0$

$= 2 - 1$

$= 1$

keep going recursivly

d) Let t be E[time for gambles stake to go down by 1]. Verify

$$t = q + p(1 + 2t)$$

Conclude gamble can expect to play forever

(3)

$t = 1 + t$

$\Rightarrow$ So t must be ∞!

Of course must verify

Same as part a

Solutions to In-Class Problems Week 14, Mon.

Problem 1.
You have a biased coin with nonzero probability $p < 1$ of coming up heads. You toss until a head comes up, and then, as in Section 18.8, you keep tossing until you get a long run of tails, but this time let "long run" mean a run of tails that is at least $k - 10$ when your initial run was length k. Prove that the expected number of times you toss a head and start over is still infinite.

Solution. Let T be the length of your initial run of tails. If $T = k$, then the expected number of tries until getting $k - 10$ tails will be the mean time to "failure," q^{k-10}, because the probability of "failing" by tossing $k - 10$ tails in a row is $q^{-(k-10)}$, where $q ::= 1 - p$. Letting R be the number of restarts, we have

$$\text{Ex}[R] = \sum_{k \in \mathbb{N}} \text{Ex}[R \mid T = k] \cdot \Pr[T = k] = \left(\sum_{k<10} q^k p \right) + \sum_{k \geq 10} \frac{1}{q^{k-10}} \cdot q^k p = \text{constant} + \sum_{k \geq 10} \frac{p}{q^{10}} = \infty.$$

■

Problem 2.
Let R be a positive integer valued random variable such that

$$\text{PDF}_R(n) = \frac{1}{cn^3},$$

where

$$c ::= \sum_{n=1}^{\infty} \frac{1}{n^3}.$$

(a) Prove that $\text{Ex}[R]$ is finite.

Solution.

$$\text{Ex}[R] ::= \sum_{n \in \mathbb{N}^+} n \cdot \frac{1}{cn^3} = \sum_{n \in \mathbb{N}^+} \frac{1}{cn^2} < 1 + \int_1^{\infty} \frac{1}{cx^2}\, dx = 1 + \frac{1}{2c} < \infty.$$

■

(b) Prove that $\text{Var}[R]$ is infinite.

Solution. Since

$$\text{Var}[R] = \text{Ex}[R^2] - \text{Ex}^2[R],$$

and $\text{Ex}^2[R] < \infty$ by part (a), we need only show that $\text{Ex}[R^2] = \infty$. But

$$\text{Ex}[R^2] ::= \sum_{n \in \mathbb{N}^+} n^2 \frac{1}{cn^3} = \sum_{n \in \mathbb{N}^+} \frac{1}{cn} = \frac{1}{c} \cdot \lim_{n \to \infty} H_n = \infty.$$

■

Problem 3.
Let T be a positive integer valued random variable such that

$$\mathrm{PDF}_T(n) = \frac{1}{an^2},$$

where

$$a ::= \sum_{n \in \mathbb{Z}^+} \frac{1}{n^2}.$$

(a) Prove that $\mathrm{Ex}[T]$ is infinite.

Solution.

$$\begin{aligned} \mathrm{Ex}[T] &::= \sum_{n \in \mathbb{Z}^+} n \mathrm{PDF}_T(n) \\ &= \sum_{n \in \mathbb{Z}^+} n \frac{1}{an^2} \\ &= \sum_{n \in \mathbb{Z}^+} \frac{1}{an} \\ &= \frac{1}{a} \lim_{n \in \mathbb{Z}^+} \cdot H_n = \infty. \end{aligned}$$

■

(b) Prove that $\mathrm{Ex}[\sqrt{T}]$ is finite.

Solution.

$$\begin{aligned} \mathrm{Ex}[\sqrt{T}] &= \sum_{n \in \mathbb{Z}^+} \sqrt{n} \cdot \frac{1}{an^2} \\ &= \sum_{n \in \mathbb{Z}^+} \frac{1}{an^{3/2}} < \int_1^\infty \frac{1}{n^{3/2}} = \frac{2}{3a}. \end{aligned}$$

■

Problem 4.
In gambler's ruin scenario, the gambler makes independent \$1 bets, where the probability of winning a bet p and of losing is $q ::= 1 - p$. The gambler keeps betting until he goes broke or reaches a target of T dollars.

Suppose $T = \infty$, that is, the gambler keeps playing until he goes broke. Let r be the probability that starting with $n > 0$ dollars, the gambler's stake ever gets reduced to $n - 1$ dollars.

(a) Explain why

$$r = q + pr^2.$$

Solution. By Total Probability

$$\begin{aligned} r &= \Pr\left[\text{ever down \$1} \mid \text{lose the first bet}\right] \Pr[\text{lose the first bet}] + \\ &\quad \Pr\left[\text{ever down \$1} \mid \text{win the first bet}\right] \Pr[\text{win the first bet}] \\ &= q + p \Pr\left[\text{ever down \$1} \mid \text{win the first bet}\right] \end{aligned}$$

But

$$
\begin{aligned}
&\Pr\left[\text{ever down \$1} \mid \text{win the first bet}\right] \\
&= \Pr[\text{ever down \$2}] \\
&= \Pr[\text{being down the first \$1}] \Pr[\text{being down another \$1}] \\
&= r^2.
\end{aligned}
$$

■

(b) Conclude that if $p \leq 1/2$, then $r = 1$.

Solution. $pr^2 - r + q$ has roots q/p and 1. So $r = 1$ or $r = q/p$. But $r \leq 1$, which implies $r = 1$ when $q/p \geq 1$, that is, when $p \leq 1/2$.

In fact $r = q/p$ when $q/p < 1$, namely, when $p > 1/2$, but this requires an additional argument that we omit.

■

(c) Conclude that even in a fair game, the gambler is sure to get ruined *no matter how much money he starts with*!

Hint: If r_n is probability of ruin starting with stake n, then $r_n = r_{n+1}p + r_{n-1}q$, so

$$r_{n+1} = \frac{r_n}{p} - r_{n-1}\frac{q}{p}. \tag{1}$$

Solution. The gambler gets ruined starting with initial stake $n = 1$ precisely if his initial stake goes down by 1 dollar, so $r_1 = r$ and $r = 1$ in the fair case. Also $r_0 = 1$ by definition. Assuming by strong induction that $r_n = r_{n-1} = 1$, the recurrence (1) implies that $r_{n+1} = 1/p - (1-p)/p = p/p = 1$. So $r_n = 1$ for all $n \geq 0$ by strong induction. ■

(d) Let t be the expected time for the gambler's stake to go down by 1 dollar. Verify that

$$t = q + p(1 + 2t).$$

Conclude that starting with a 1 dollar stake in a fair game, the gambler can expect to play forever!

Solution. By Total Expectation

$$
\begin{aligned}
t &= \text{Ex}[\text{\#steps to be down \$1} \mid \text{lose the first bet}] \Pr[\text{lose the first bet}] + \\
&\qquad \text{Ex}[\text{\#steps to be down \$1} \mid \text{win the first bet}] \Pr[\text{win the first bet}] \\
&= q + p\,\text{Ex}[1 + \text{\#steps to be down \$1} \mid \text{win the first bet}].
\end{aligned}
$$

But

$$
\begin{aligned}
&\text{Ex}[\text{\#steps to be down \$1} \mid \text{win the first bet}] \\
&= \text{Ex}[\text{\#steps to be down \$2}] \\
&= \text{Ex}[\text{\#steps to be down the first \$1}] + \text{Ex}[\text{\#steps to be down another \$1}] \\
&= 2t.
\end{aligned}
$$

This implies the required formula $t = q + p(1 + 2t)$. If $p = 1/2$ we conclude that $t = 1 + t$, which means t must be infinite.

■

5/11

Mathematics for Computer Science

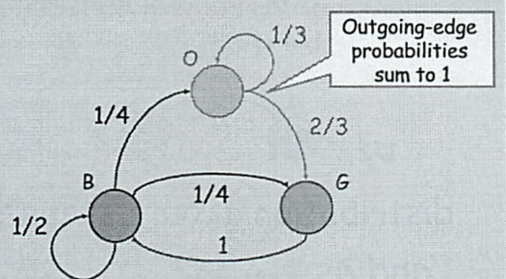
MIT 6.042J/18.062J

Random Walks

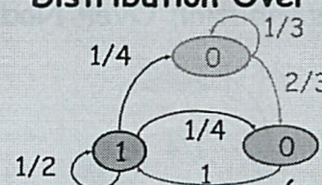
Applications of Random Walk

- Physics — Brownian motion
- Finance — stocks, options
- Algorithms — web search, clustering

Graph With Probable Transitions



Distribution Over Nodes

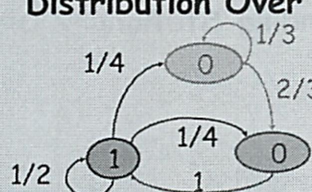


(p_B, p_O, p_G)

Suppose you start at B: $(1, 0, 0)$

What are p'_B, p'_O, p'_G after 1 step?

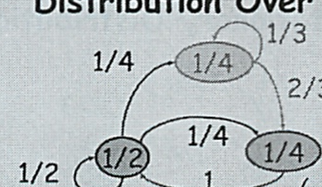
Distribution Over Nodes



Dist after 1 step: (p'_B, p'_G, p'_O)

only get places from B, so $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$

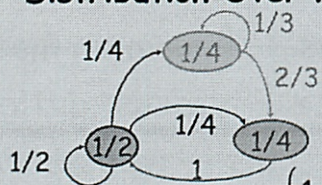
Distribution Over Nodes



Dist after 1 step: $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$

Dist after 2 steps: (p''_B, p''_O, p''_G)

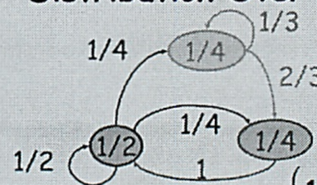
Distribution Over Nodes



Dist after 1 step: $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$

$p''_O = \Pr\{B \text{ to } O \mid \text{at } B\} \cdot p'_B$
$+ \Pr\{O \text{ to } O \mid \text{at } O\} \cdot p'_O$
$+ \Pr\{G \text{ to } O \mid \text{at } G\} \cdot p'_G$

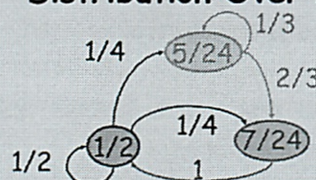
Distribution Over Nodes



Dist after 1 step: $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$

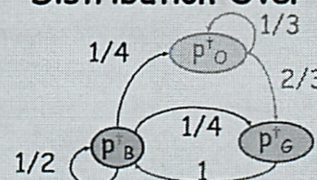
$$p''_O = \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{4} + 0 \cdot \frac{1}{4} = 5/24$$

Distribution Over Nodes



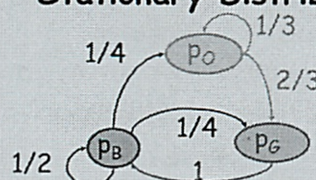
distribution after 2 steps:
(p''_B, p''_O, p''_G)
$\left(\frac{1}{2}, \frac{5}{24}, \frac{7}{24}\right)$

Distribution Over Nodes



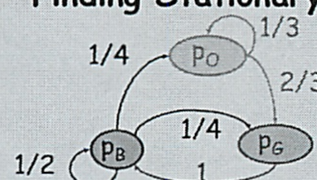
distribution after t steps?
...and as $t \to \infty$?

Stationary Distribution



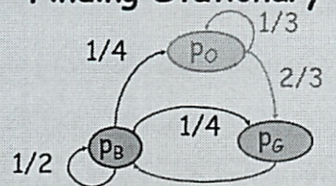
distribution (p_B, p_O, p_G) is stationary if next-step distribution is the same.
What is a stationary dist. here?

Finding Stationary Dist.



$p_B = p_B' = (1/2)p_B + 1p_G$
$p_O = p_O' = (1/4)p_B + (1/3)p_O$
$p_G = p_G' = (1/4)p_B + (2/3)p_O$
$p_B + p_O + p_G = 1$

Finding Stationary Dist.



solving for (p_B, p_O, p_G): $\left(\frac{8}{15}, \frac{3}{15}, \frac{4}{15}\right)$

Google Page Rank

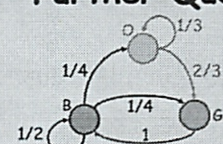
View the entire web as a graph
- vertices are webpages
- edge (u,v) exists if link from page u to page v
- Pr{go to v from u} = 1/outdeg(u)

Find stationary distribution $\{p_u\}$
Rank u above v if $p_u > p_v$.

Questions on Stationary Dist

- Does a stationary dist exist? Yes (if graph finite)
- Is it unique? Sometimes
- Does a random walk approach it from any starting distribution? Sometimes
  - How quickly? Varies

Further Questions



- Pr{ever reach O | start at B}
- Pr{reach G before O | start at B}
- Average # steps for B to O

Team Problems

Problems

1 -- 3

6.042 Random Walks

5/11

- Filled out evals

- Physics
- Finance
- CS : Google

Graphs w/ probable transitions
- Outgoing edge probabilities sum to 1

Given a certain start state, what is prob end up somewhere?
- After a bunch of time steps
- Can calculate step by step
  - Add up all the possible inputs
- But how to do t steps
  - 3×3 multiplication
  - $\log t$ squarings
- What happens as $t \to \infty$?

Needs to be stationary
- or else would not approach a limit

②

Solve a system of linear eqn

~~Want 2 do~~ Solve for items in terms of other items

(see slides)

(Same as 6.041 Markov Chains)

Google Page Rank works like this

- How to sort pages?
- Each page is node ~~as~~ of a graph
- links in & out
- large # links in - need to come from important sites

This is the core idea - Google has 100 people improving

$$p(v \text{ from } u) = \frac{1}{\text{outdeg}(v)}$$

Does a stationary dist exist?

- If cycle - no
- Need no ~~dea~~ vertex w/ degout = 0

So Google added supervertex
- Makes graph strongly connected

③

~~[illegible]~~

How fast does it converge?

Can ask

- will you ever reach
- reach one before other

3 coins

- can think of as ~~last~~ 8 state graph

Or one coin where last heads matter

Final Format

12× 15 min ~~[illegible]~~ qu

9 of 12 will be small pertibutions of previous qu

3 of 12 integrates various sections

Class Grades

20-30% get A- and above

65-75% B-B-

~~the rest~~ get D, F ~~5-15%~~
5-15%

The rest Cs

7-faces of cheat sheets for final

In-Class Problems Week 14, Wed.

Problem 1. **(a)** Find a stationary distribution for the random walk graph in Figure 1.

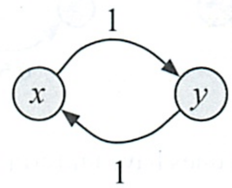


Figure 1

(b) If you start at node x in Figure 1 and take a (long) random walk, does the distribution over nodes ever get close to the stationary distribution? Explain.

(c) Find a stationary distribution for the random walk graph in Figure 2.

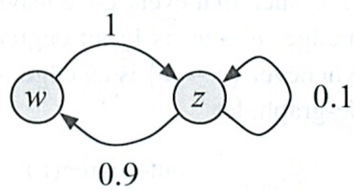


Figure 2

(d) If you start at node w Figure 2 and take a (long) random walk, does the distribution over nodes ever get close to the stationary distribution? You needn't prove anything here, just write out a few steps and see what's happening.

(e) Find a stationary distribution for the random walk graph in Figure 3.

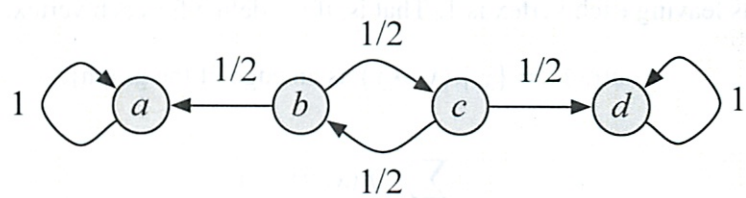


Figure 3

(f) If you start at node b in Figure 3 and take a long random walk, the probability you are at node d will be close to what fraction? Explain.

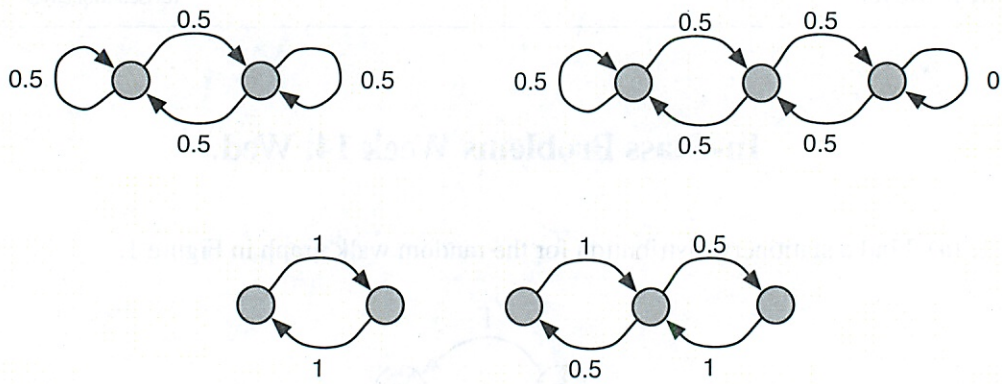


Figure 4 Which ones have uniform stationary distribution?

Problem 2.
For which of the graphs in Figure 4 is the uniform distribution over nodes a stationary distribution? The edges are labeled with transition probabilities. Explain your reasoning.

Problem 3.
A Google-graph is a random-walk graph such that every edge leaving any given vertex has the same probability. That is, the probability of each edge $\langle v \to w \rangle$ is $1/\text{out-degree}(v)$.

A directed graph is *symmetric* if, whenever $\langle v \to w \rangle$ is an edge, so is $\langle w \to v \rangle$.

Given any finite, symmetric Google-graph, let

$$d(v) ::= \frac{\text{out-degree}(v)}{e},$$

where e is the total number of edges in the graph. Show that d is a stationary distribution.

Appendix

A *random-walk graph* is a digraph such that each edge, $\langle x \to y \rangle$, is labelled with a number, $p(x, y) > 0$, which will indicate the probability of following that edge starting at vertex x. Formally, we simply require that the sum of labels leaving each vertex is 1. That is, if we define for each vertex, x,

$$\text{out}(x) ::= \{y \mid \langle x \to y \rangle \text{ is an edge of the graph}\},$$

then

$$\sum_{y \in \text{out}(x)} p(x, y) = 1.$$

A *distribution*, d, is a labelling of each vertex, x, with a number, $d(x) \geq 0$, which will indicate the probability of being at x. Formally, we simply require that the sum of all the vertex labels is 1, that is,

$$\sum_{x \in V} d(x) = 1,$$

where V is the set of vertices.

The distribution, $\widehat{d}$, *after a single step* of a random walk from distribution, d, is given by

$$\widehat{d}(x) ::= \sum_{y \in \text{in}(x)} d(y) \cdot p(y, x),$$

where

$$\text{in}(x) ::= \{y \mid \langle y \to x \rangle \text{ is an edge of the graph}\}.$$

A distribution d is *stationary* if $\widehat{d} = d$, where $\widehat{d}$ is the distribution after a single step of a random walk starting from d. In other words, d stationary implies

$$d(x) ::= \sum_{y \in \text{in}(x)} d(y) \cdot p(y, x).$$

In-Class 14 Wed

5/11

1.a) $x \to \frac{1}{2}$

$y \to \frac{1}{2}$

b) ~~Yes - it does not change~~

~~Does not~~

No - I remember where you are is strictly defined by where you start

* even time step x
odd " " y

c) Here is different. What is formula again?

$$p_w = .9\, p_z$$

$$p_z = p_w + .1\, p_z$$

$$p_w + p_z = 1$$

- solve

$\cancel{p_z = .9 p_z + .1 p_z}$ Not interesting

$$p_z = p_w + .1 \frac{p_w}{.9} = \frac{10}{9} p_w$$

② Still not right - what am I doing wrong?

Use $p_w + p_z = 1$

$p_w = .9(1 - p_w)$

$p_w = .9 - .9p_w$

$1.9p_w = .9$

$p_w = .47368$ ← there we go

$p_z = 1 - \uparrow = .526$

d) Yes - since random now

?what is long explination they are looking for?

e) ~~para~~ It will end up at a, d each w/ ~~M~~

$P() = \frac{1}{2}$ - depending on where it starts (b or c)

f) $\frac{1}{2}$ - will fall in either hole/sink w/ equal probability

Q3

2. Which has uniform dist?

top ~~right~~ left ? top right

not bottom left not bottom right

Just intuition - I would verify on an exam

3. Google-graph - random walk so every ~~the~~ edge has = prob.

$$\frac{1}{out(v)}$$

~~Graph~~

$$d(v) = \frac{out(v)}{e}$$ ← total # edges of graph

Show d is stat dist

? So the long-term markov chain

w/ outgoing prob of each link restricted $\frac{1}{out(v)}$

? Wouldn't it be in degree?

9

2 board) ~~they~~ They did verify it

$\frac{1}{2}\left(\frac{1}{2}\right) + \frac{1}{2}\left(\frac{1}{2}\right) = \frac{1}{2}$

$\frac{1}{2}\left(\frac{1}{2}\right) + \frac{1}{2}\left(\frac{1}{2}\right) = \frac{1}{2}$ uniform

$\frac{1}{2}\left(\frac{1}{3}\right) + \frac{1}{2}\left(\frac{1}{3}\right) = \frac{1}{3}$ uniform

" " "

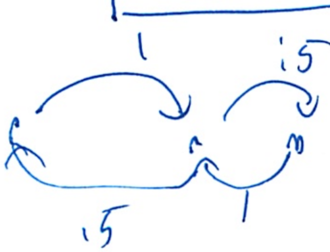
" " "

$1\left(\frac{1}{2}\right) = \frac{1}{2}$ uniform

$1\left(\frac{1}{2}\right) = \frac{1}{2}$

hmm, why (see 1b)

it will never get there but it still have that prob



No

$\frac{1}{2}\left(\frac{1}{3}\right) = \frac{1}{6}$

$1\left(\frac{1}{3}\right) + 1\left(\frac{1}{3}\right) = \frac{2}{3}$

$\frac{1}{2}\left(\frac{1}{3}\right) = \frac{1}{6}$

(5)

3 are bored | ~~[illegible]~~

$P_v = d(v)$ is stationary dist iff $p_v' = P_v = d(v)$

Assuming $P_v = d(v)$ $\forall$ vertices v in symetric Google graph

For any vertex v,

$$P_v' = \sum_{w \in N(v)} P(\text{transition to } v / w) P(w)$$

$$= \sum_{w \in N(v)} \frac{1}{out(v)} * \frac{out\ deg(v)}{e}$$

$$= \sum_{w \in N(v)} \frac{1}{e}$$

$$= \frac{indeg(v)}{e}$$

$indeg(v) = out\ deg(v)$ in symetricg graph

?where did they say it was symetric

$P_v' = \frac{out(v)}{e} = d(v) = P_v$

So $P_v = d(v)$ is stationary

Solutions to In-Class Problems Week 14, Wed.

Problem 1. **(a)** Find a stationary distribution for the random walk graph in Figure 1.

Solution. $d(x) = d(y) = 1/2$ ■

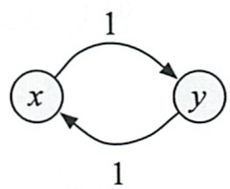


Figure 1

(b) If you start at node x in Figure 1 and take a (long) random walk, does the distribution over nodes ever get close to the stationary distribution? Explain.

Solution. No! you just alternate between nodes x and y. ■

(c) Find a stationary distribution for the random walk graph in Figure 2.

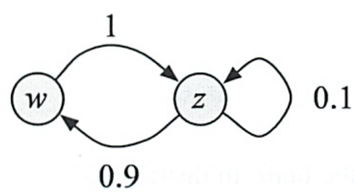


Figure 2

Solution. $d(w) = 9/19$, $d(z) = 10/19$. You can derive this by setting $d(w) = (9/10)d(z)$, $d(z) = d(w) + (1/10)d(z)$, and $d(w) + d(z) = 1$. There is a unique solution. ■

(d) If you start at node w Figure 2 and take a (long) random walk, does the distribution over nodes ever get close to the stationary distribution? You needn't prove anything here, just write out a few steps and see what's happening.

Solution. Yes, it does. ■

(e) Find a stationary distribution for the random walk graph in Figure 3.

Solution. There are infinitely many, with $d(b) = d(c) = 0$, and $d(a) = p$ and $d(d) = 1 - p$ for any p. ■

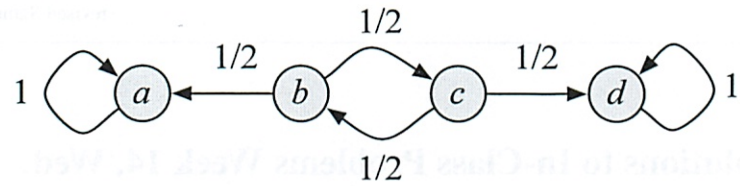


Figure 3

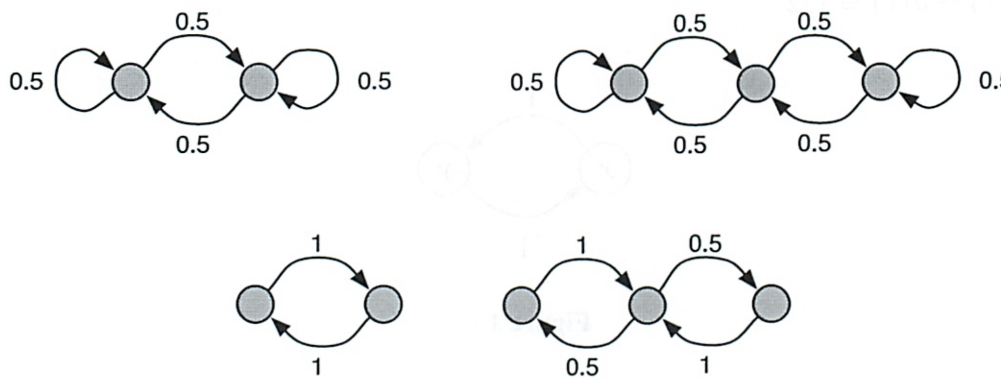


Figure 4 Which ones have uniform stationary distribution?

(f) If you start at node b in Figure 3 and take a long random walk, the probability you are at node d will be close to what fraction? Explain.

Solution. 1/3. ■

Problem 2.
For which of the graphs in Figure 4 is the uniform distribution over nodes a stationary distribution? The edges are labeled with transition probabilities. Explain your reasoning.

Solution. All except the last one (bottom right).

One way of approaching this problem is by performing a single update step according to the rule

$$\widehat{d}(v) = \sum_{u \text{ s.t. } \langle u \to v \rangle} d(u)p(u,v),$$

where d is the stationary distribution ($1/2$ for all vertices on the left graphs, $1/3$ for all vertices on the right), $\widehat{d}$ is the distribution after one step, and $p(u,v)$ is the edge probability. If $\widehat{d} = d$, then by definition, the uniform distribution is stationary.

Alternatively, you could observe that the uniform distribution is stationary if and only if $\widehat{d}(v) = d(v)$, and hence dividing both sides by probability of being at each vertex, we get

$$1 = \sum_{u \text{ s.t. } \langle u \to v \rangle} p(u,v).$$

In other words, the uniform distribution is stationary if and only if the incoming-edge probabilities sum to 1. ■

Problem 3.
A Google-graph is a random-walk graph such that every edge leaving any given vertex has the same probability. That is, the probability of each edge $\langle v \to w \rangle$ is $1/\text{out-degree}(v)$.

A directed graph is *symmetric* if, whenever $\langle v \to w \rangle$ is an edge, so is $\langle w \to v \rangle$.

Given any finite, symmetric Google-graph, let

$$d(v) ::= \frac{\text{out-degree}(v)}{e},$$

where e is the total number of edges in the graph. Show that d is a stationary distribution.

Solution. To show that d is a stationary distribution, we must show that

$$d(w) = \sum_{v \in \text{in}(w)} p(v, w) d(v), \tag{1}$$

where $\text{in}(w) ::= \{v \mid \langle v \to w \rangle \text{ is an edge}\}$.

We have

$$\begin{aligned}
&\sum_{v \in \text{in}(w)} p(v, w) d(v) \\
&= \sum_{v \in \text{in}(w)} \left(\frac{1}{\text{out-degree}(v)}\right)\left(\frac{\text{out-degree}(v)}{e}\right) && \text{(by def } p \text{ and } d\text{)} \\
&= \sum_{v \in \text{in}(w)} \frac{1}{e} \\
&= |\text{in}(w)| \frac{1}{e} \\
&= \text{in-degree}(w) \frac{1}{e} \\
&= \text{out-degree}(w) \frac{1}{e} && \text{(by symmetry of the graph)} \\
&= d(w)\ .
\end{aligned}$$

■

Appendix

A *random-walk graph* is a digraph such that each edge, $\langle x \to y \rangle$, is labelled with a number, $p(x, y) > 0$, which will indicate the probability of following that edge starting at vertex x. Formally, we simply require that the sum of labels leaving each vertex is 1. That is, if we define for each vertex, x,

$$\text{out}(x) ::= \{y \mid \langle x \to y \rangle \text{ is an edge of the graph}\},$$

then

$$\sum_{y \in \text{out}(x)} p(x, y) = 1.$$

A *distribution*, d, is a labelling of each vertex, x, with a number, $d(x) \geq 0$, which will indicate the probability of being at x. Formally, we simply require that the sum of all the vertex labels is 1, that is,

$$\sum_{x \in V} d(x) = 1,$$

where V is the set of vertices.

The distribution, $\widehat{d}$, *after a single step* of a random walk from distribution, d, is given by

$$\widehat{d}(x) ::= \sum_{y \in \text{in}(x)} d(y) \cdot p(y, x),$$

where

$$\text{in}(x) ::= \{y \mid \langle y \to x \rangle \text{ is an edge of the graph}\}.$$

A distribution d is *stationary* if $\widehat{d} = d$, where $\widehat{d}$ is the distribution after a single step of a random walk starting from d. In other words, d stationary implies

$$d(x) ::= \sum_{y \in \text{in}(x)} d(y) \cdot p(y, x).$$