Expected value over countably infinite sets

$S =$ a countably infinite subset of R

$S = \{s_1, s_2, \ldots\}$

$X =$ a random variable which gets values in S

$$E(X) \doteq \sum_{i=1}^{\infty} s_i \Pr(X(\omega) = s_i)$$

Consider the distribution

distribution is over the natural numbers = positive integers

$$P(X = i) = \frac{1}{zi^3}; \quad z = \sum_{i=1}^{\infty} \frac{1}{i^3} < \infty$$

$$E(X) = \sum_{i=1}^{\infty} \frac{i}{zi^3} = \sum_{i=1}^{\infty} \frac{1}{zi^2} < \infty$$

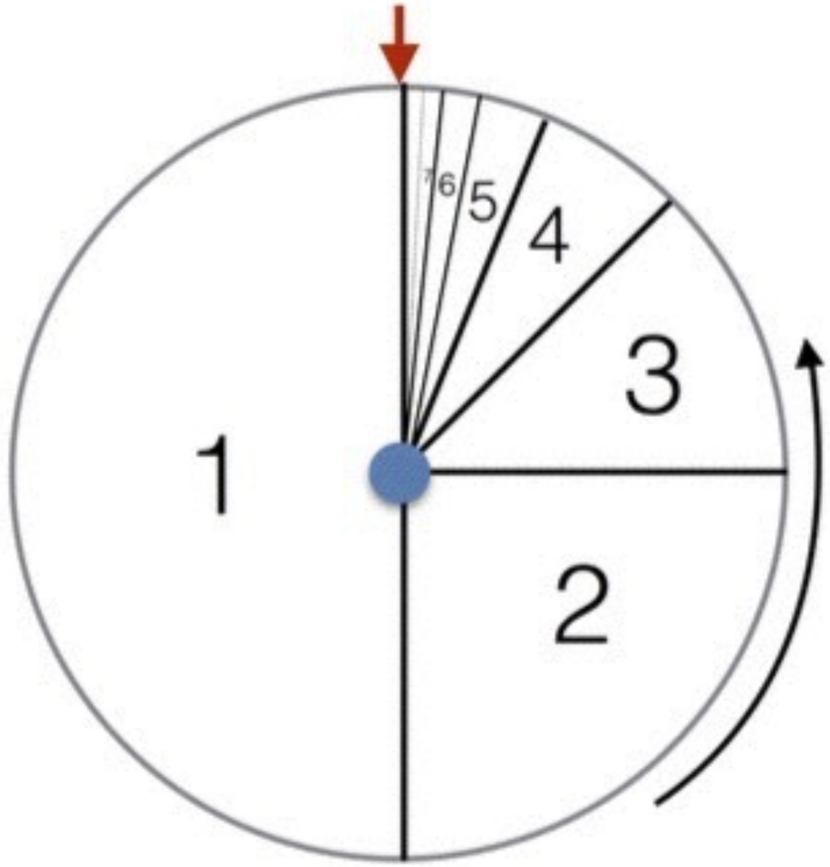
Expectation is finite

Consider next the distribution

$$P(X = i) = \frac{1}{zi^2}; \quad z = \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty$$

$$E(X) = \sum_{i=1}^{\infty} \frac{i}{zi^2} = \sum_{i=1}^{\infty} \frac{1}{zi} = \infty$$

Distribution is well defined but Expectation is infinite



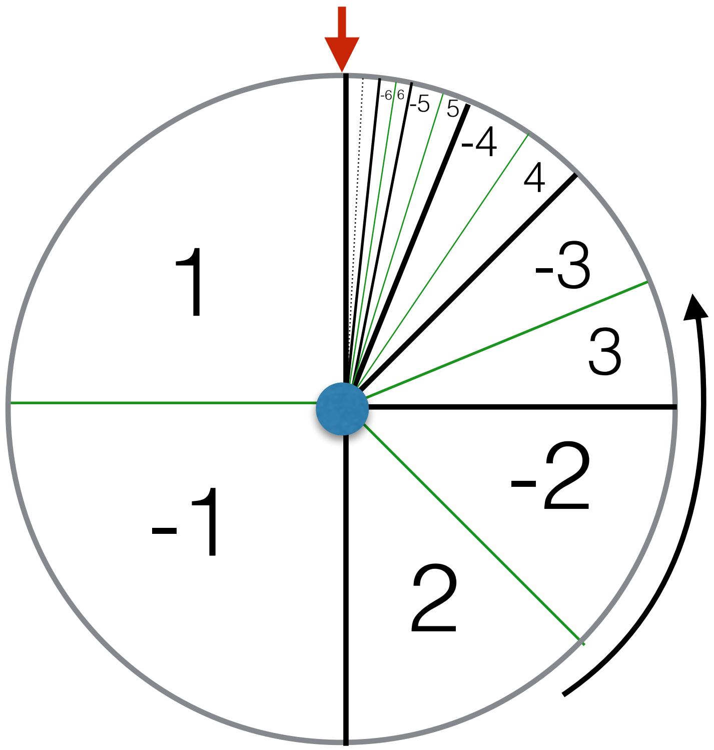
(b) A wheel with
Infinitely many outcomes

$$P(X = i) = \frac{6}{\pi^2 i^2};$$

$$\sum_{i=1}^{\infty} P(X = i) = 1$$

$$\sum_{i=1}^{\infty} iP(X = i) = \infty$$

Participation in this game is worth any price
(on the long term)



A wheel with
Infinitely many outcomes
both positive and negative

Consider a game with both wins and losses

$$i \in \{0,-1,+1,-2,+2,\ldots\}$$

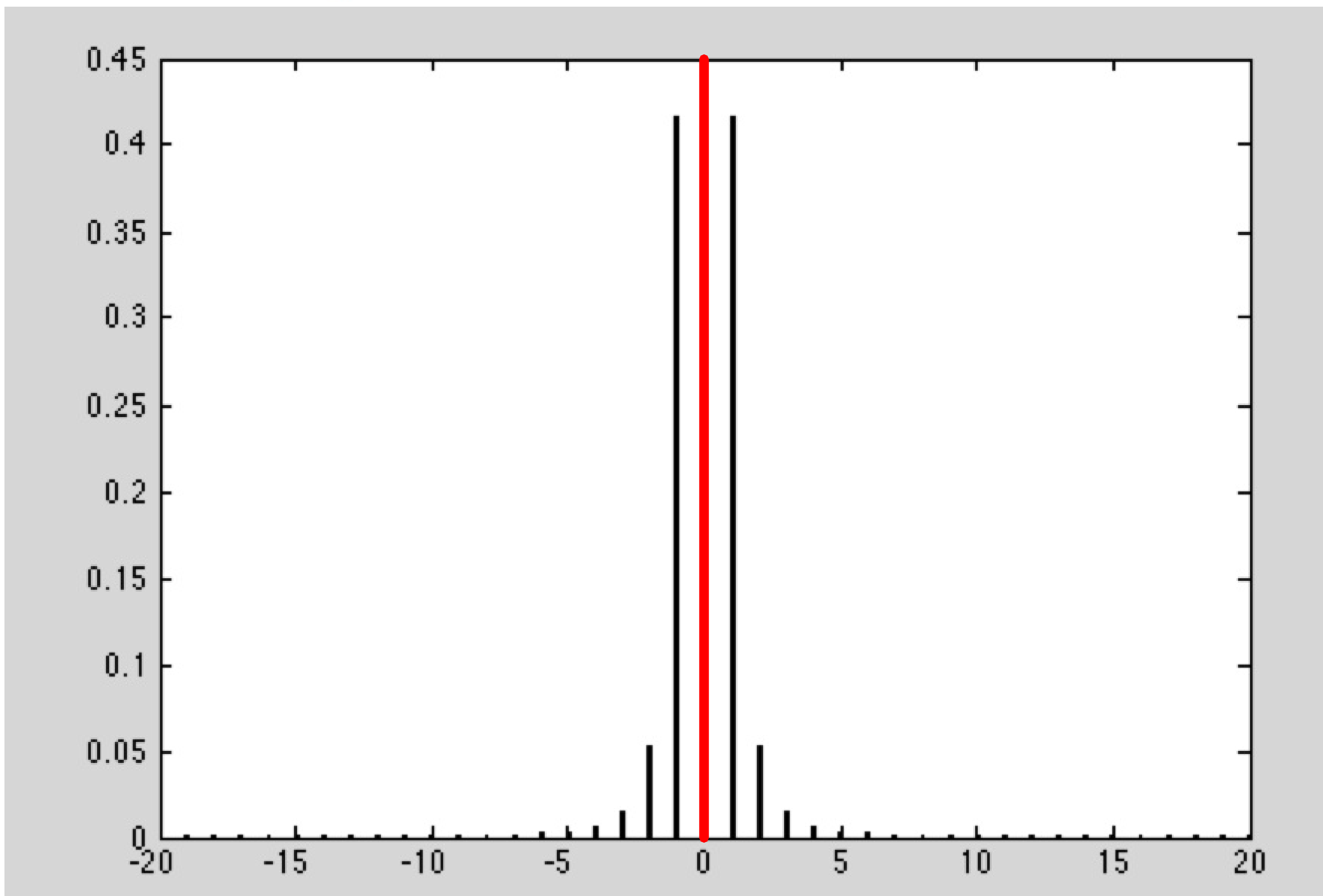
$$P(X=i)=\begin{cases} \dfrac{1}{Z}\dfrac{1}{i^{1.5}} & if \quad i \neq 0 \\ 0 & if \quad i=0 \end{cases} \quad , \quad Z=2\sum_{i=1}^{\infty}\frac{1}{i^{1.5}}$$

$$\sum_{i=-\infty}^{\infty} P(X=i)=1$$

$$\sum_{i=\infty}^{\infty} iP(X=i) \;\; is\ undefined$$

Expectation over pos and neg integers: the good case

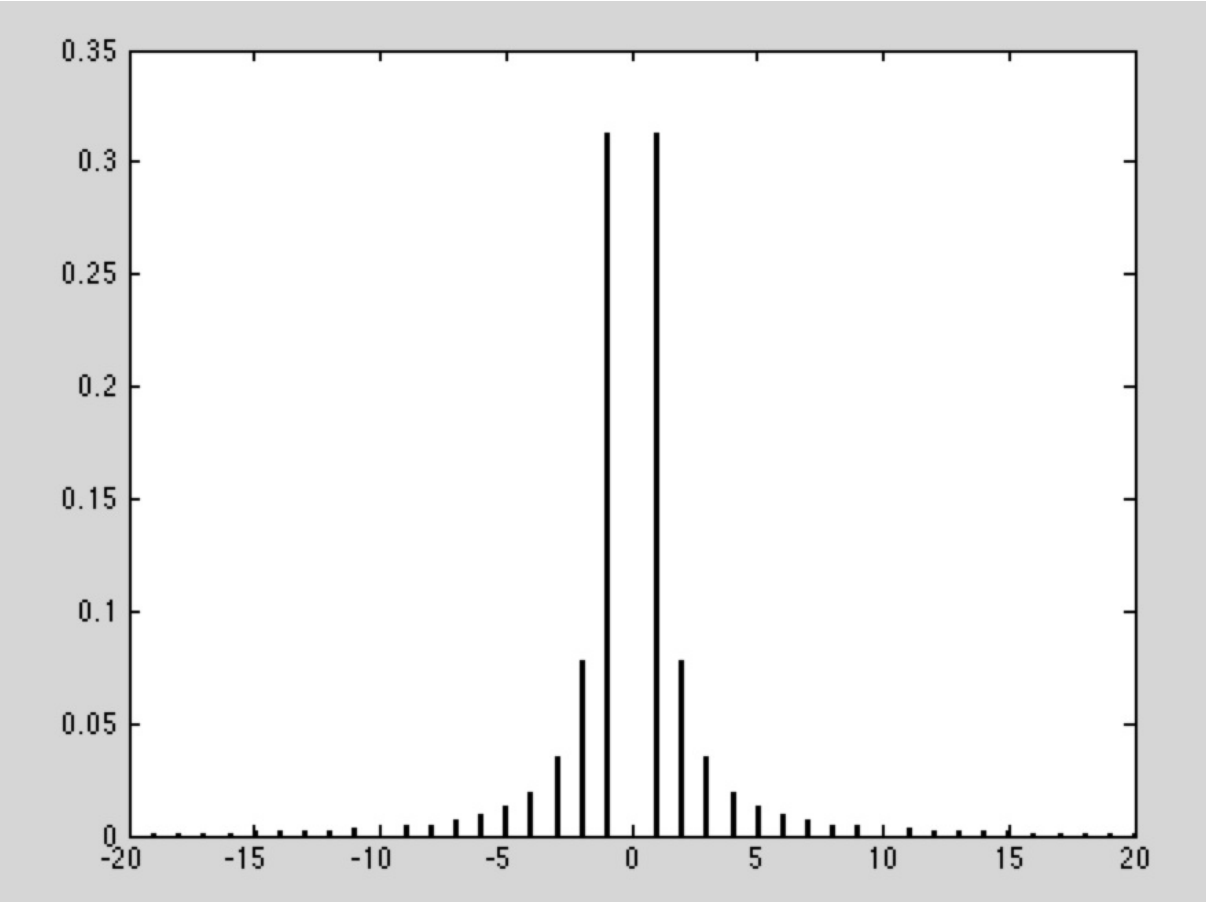
$$P(X=i)=\begin{cases} 0 & \text{if } i=0 \\ \dfrac{1}{Z|i|^3} & \text{if } i \neq 0 \end{cases} \quad ; \quad Z = 2\sum_{i=1}^{\infty}\frac{1}{|i|^3} < \infty$$



$$E(X)=\sum_{i=1}^{\infty} iP(X=i)+\sum_{i=-1}^{-\infty} iP(X=i)=\frac{1}{Z}\left(\sum_{i=1}^{\infty}\frac{1}{i^2}-\sum_{i=1}^{\infty}\frac{1}{i^2}\right)=\frac{c-c}{Z}=0$$

A symmetric distribution on pos and neg integers, the bad case

$$P(X=i)=\begin{cases} 0 & \text{if } i=0 \\ \dfrac{1}{Zi^2} & \text{if } i\neq 0 \end{cases} \quad ; \quad Z=2\sum_{i=1}^{\infty}\frac{1}{i^2}<\infty$$

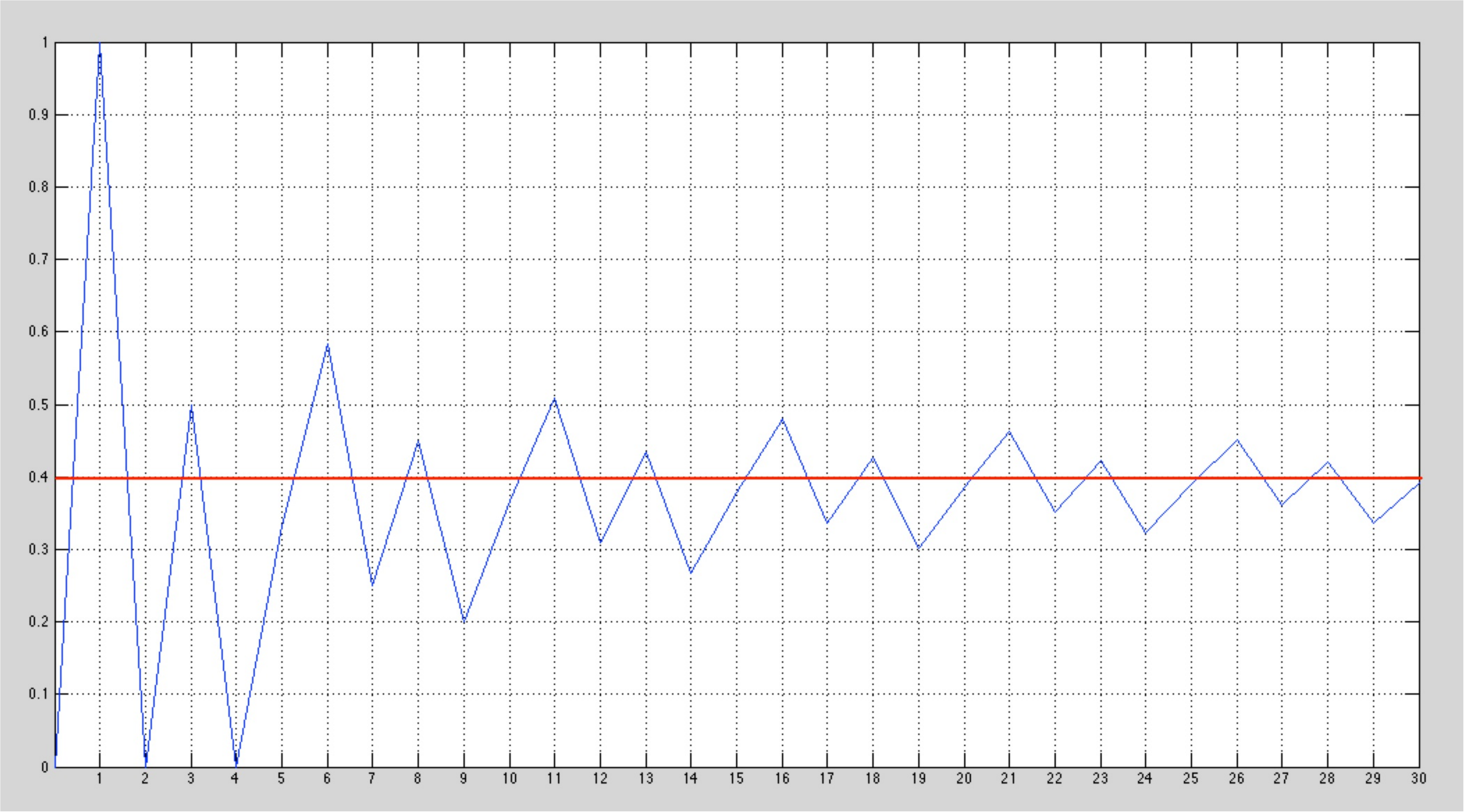


$$E(X)=\sum_{i=1}^{\infty} iP(X=i)+\sum_{i=-1}^{-\infty} iP(X=i)=\frac{1}{Z}\left(\sum_{i=1}^{\infty}\frac{1}{i}-\sum_{i=1}^{\infty}\frac{1}{i}\right)=\frac{\infty-\infty}{Z}=\mathit{undefined}$$

Undefined limit means you can get the limit of your choice by changing the order of summation.

You have at your disposal two infinitely large sums with shrinkingly small pieces:
1/1,1/2,1/3,1/4,....   -1/1,-1/2,-1/3,-1/4,...
Suppose you want the limit to be 0.4, by alternating between positives and negatives you can get arbitrarily close to 0.4 (or to any other number)

1/1-1/1+1/2-1/2+1/3+1/4-1/3+1/5-1/4+1/6+1/7-1/5+1/8-1/6+1/9+1/10-1/7
+1/11-1/8+1/12+1/13-1/9+1/14-1/10+1/15+1/16-1/11+1/17-1/12+1/18 = 0.3919



webwork.cse.ucsd.edu/misc/expected.html

Apps | Recent changes - Se | My Page | CSE103 | AWS | Googles | TWiki | UCSD

Let X be a random variable whose probability distribution is defined as:

$$P(X=i)=\begin{cases}\frac{1}{|i|^{\alpha}} & \text{if } i \neq 0\\ 0 & \text{if } i=0.\end{cases}$$

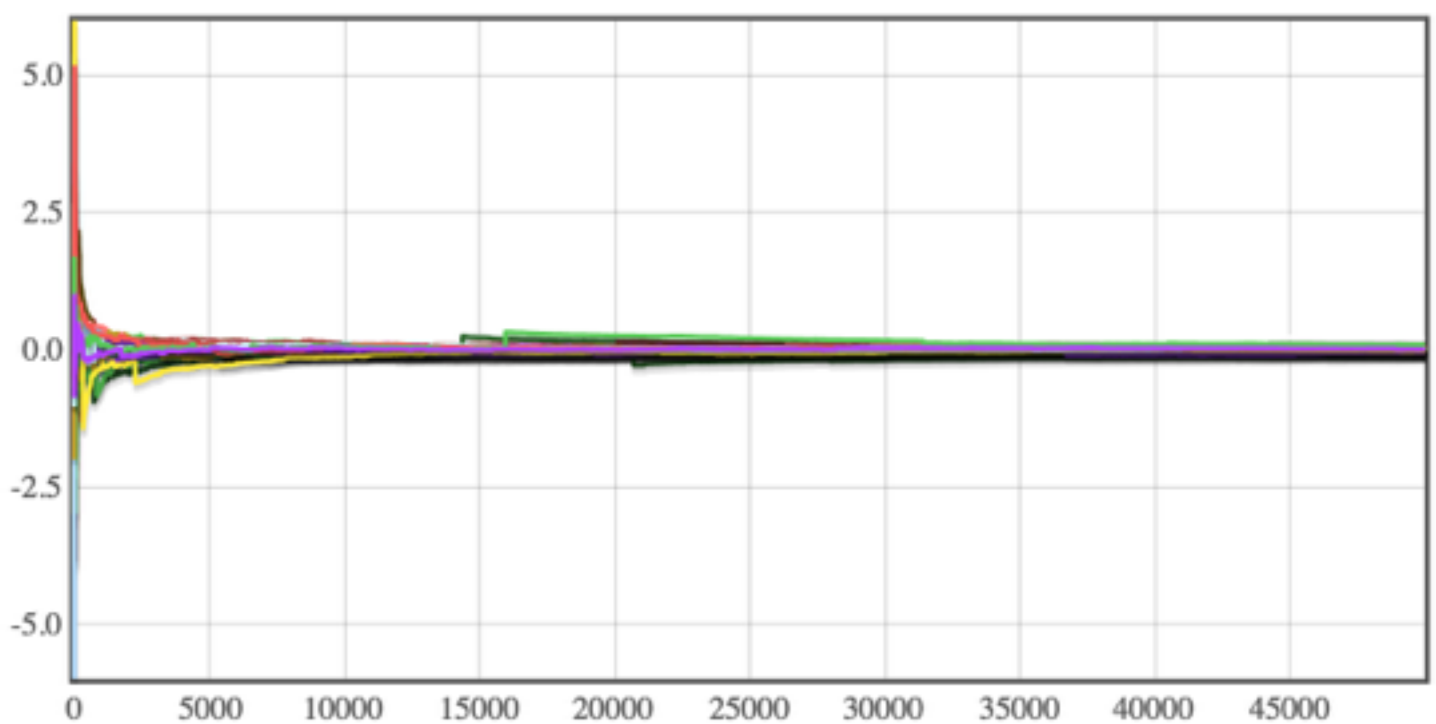
Simulation parameters:

α: 2.5

Number of trajectories: 50

Number of data points: 50000

Run



Let X be a random variable whose probability distribution is defined as:

$$P(X = i) = \begin{cases} \frac{1}{|i|^{\alpha}} & \text{if } i \neq 0 \\ 0 & \text{if } i = 0. \end{cases}$$

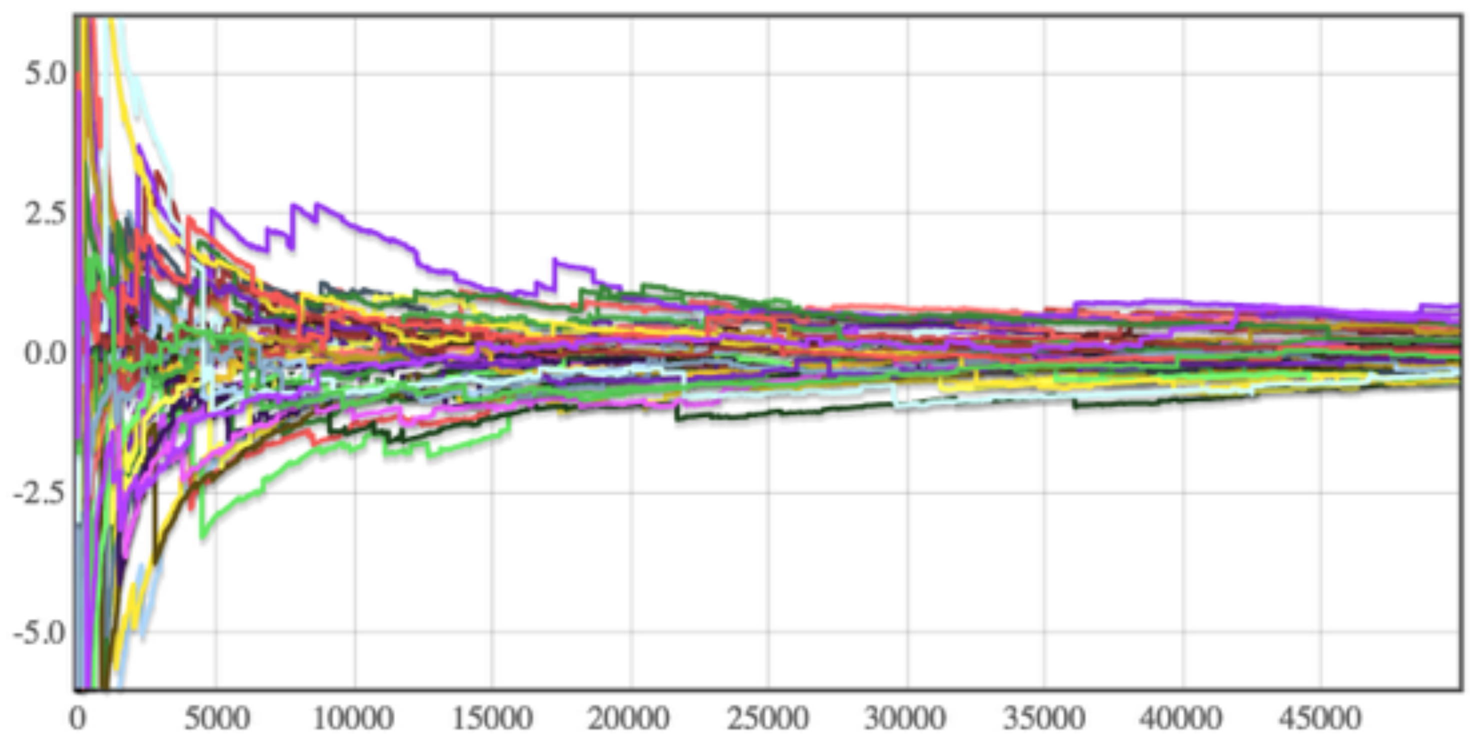
Simulation parameters:

α: 2.0

Number of trajectories: 50

Number of data points: 50000

Run



webwork.cse.ucsd.edu/misc/expected.html

Apps | Recent changes - Se | My Page | CSE103 | AWS | Googles | TWiki | UCSD

Let X be a random variable whose probability distribution is defined as:

$$P(X=i) = \begin{cases} \frac{1}{|i|^{\alpha}} & \text{if } i \neq 0 \\ 0 & \text{if } i = 0. \end{cases}$$

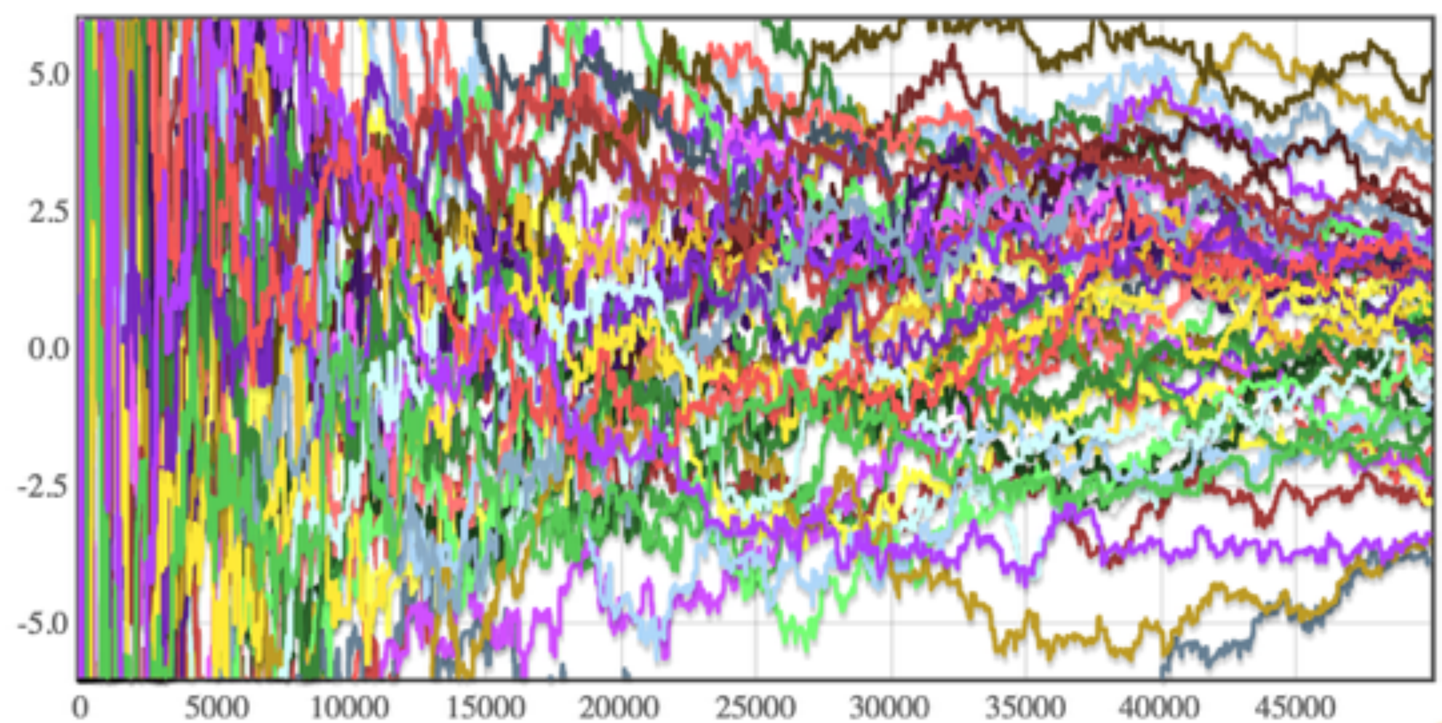
Simulation parameters:

α: 1.5

Number of trajectories: 50

Number of data points: 50000

Run



KQS cards

I want to know how to improve my teaching
Your feedback is important to me.

On the index card provided, please give a one-sentence to each of the following questions

Keep: One thing I should keep doing?
Quit: One thing I should quit doing?
Start: One thing I should start doing?

A random algorithm for computing percentiles

A problem with the average

$$Average(X_1,\ldots,X_n) \doteq \frac{1}{n}\sum_{i=1}^{n} X_i$$

The average is the most common estimator of the "center" of a distribution.
It takes linear time to compute.

However, the average is "sensitive to outliers" :

Suppose that you have a company in which 1000 employees earn 1$/day and one employee earns 1000$/day. The average dayly pay is 2000/1001 ~ 2$/day, but that is double what most people earn.

Using the Median instead of the average

To compute the median sort all n elements from smallest to largest and take the value of the element that is the middle of the list (position n/2) (take the average of the two middle elements if the list length is even).

In the earlier example, the median will be 1$ regardless of how big is the largest salary - outliers are ignored.

The median is a special case of percentiles

To compute the P-percentile sort all n elements from smallest to largest and take the value of the element that is in the floor(Pn) position.

The Median is the 1/2-percentile

Often used when describing distribution of income:
The top

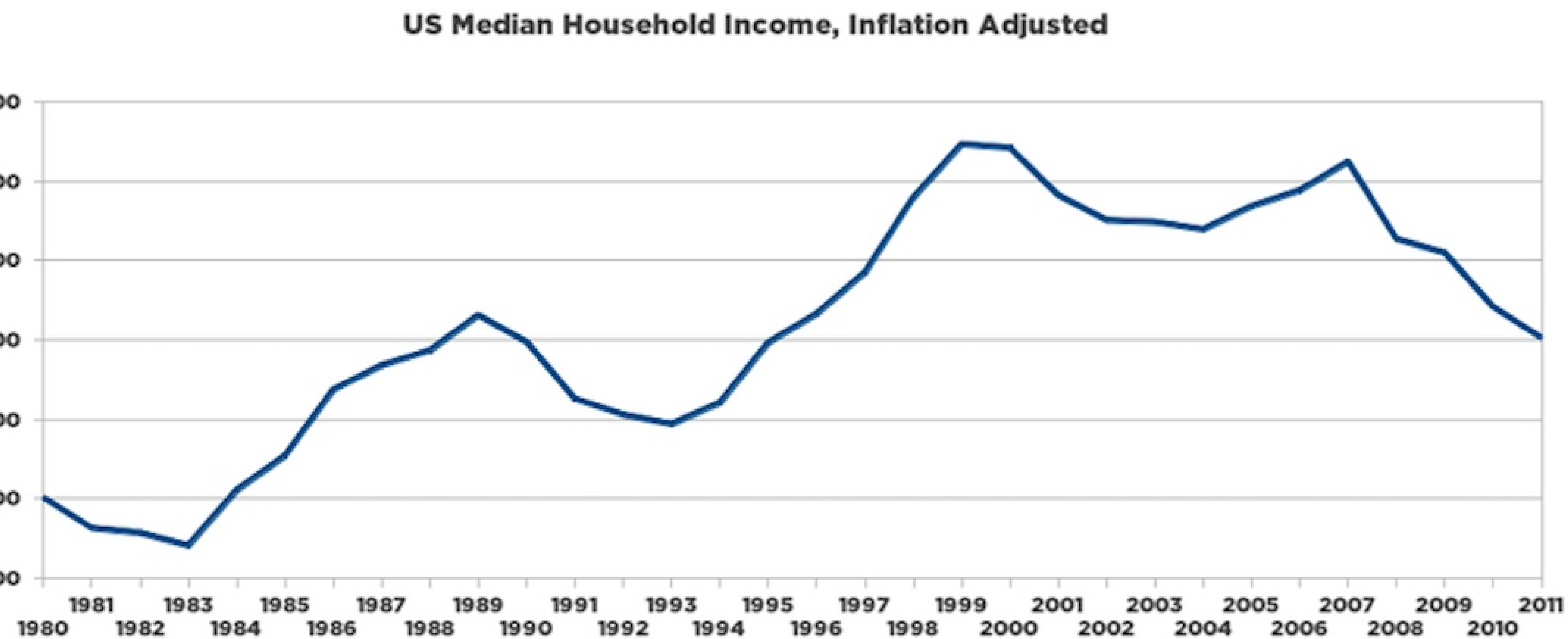

US Median Household Income, Inflation Adjusted
56,000
54,000
52,000
50,000
48,000
46,000
44,000
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011

Table 1: Income, net worth, and financial worth in the U.S. by percentile, in 2010 dollars

Wealth or income class	Mean household income	Mean household net worth	Mean household financial (non-home) wealth
Top 1 percent	$1,318,200	$16,439,400	$15,171,600
Top 20 percent	$226,200	$2,061,600	$1,719,800
60th-80th percentile	$72,000	$216,900	$100,700
40th-60th percentile	$41,700	$61,000	$12,200
Bottom 40 percent	$17,300	-$10,600	-$14,800

From Wolff (2012); only mean figures are available, not medians. Note that income and wealth are separate measures; so, for example, the top 1% of income-earners is not exactly the same group of people as the top 1% of wealth-holders, although there is considerable overlap.

A linear time algorithm for computing percentiles

We can calculate P-percentile by sorting and then picking the element in location Pn. But this requires time O(n log n) (unless distribution of data is known).

We will now describe a randomized algorithm whose expected running time is O(n)

Find the 5th smallest element in the following table:

7	**10**	**3**	**20**	**15**	**16**	**70**	**33**	**30**	**7**

Solution 1, sort and locate ---- takes worst case time O(n log n):

3	**7**	**7**	**10**	**15**	**16**	**20**	**30**	**33**	**70**

Expected time is also Omega(n log n):

Find the 5th smallest element in the following table:

7	**10**	**3**	**20**	**15**	**16**	**70**	**33**	**30**	**7**

Solution 2, randomized algorithm ---- takes ***expected*** time O(n):

Find the 5th smallest element in the following table:

7	**10**	**3**	**20**	**15**	**16**	**70**	**33**	**30**	**7**

Choose a random element as pivot

7	**10**	**3**	**20**	**15**	**16**	**70**	**33**	**30**	**7**

Find the 5th smallest element in the following table:

7	**10**	**3**	**20**	**15**	**16**	**70**	**33**	**30**	**7**

Choose a random element as pivot

7	**10**	**3**	**20**	**15**	**16**	**70**	**33**	**30**	**7**

Partition list into 3 lists: <7 , =7 , >7

Find the 5th smallest element in the following table:

7	**10**	**3**	**20**	**15**	**16**	**70**	**33**	**30**	**7**

Choose a random element as pivot

7	**10**	**3**	**20**	**15**	**16**	**70**	**33**	**30**	**7**

Partition list into 3 lists: <7 , =7 , >7

3

S_L

Find the 5th smallest element in the following table:

7	**10**	**3**	**20**	**15**	**16**	**70**	**33**	**30**	**7**

Choose a random element as pivot

7	**10**	**3**	**20**	**15**	**16**	**70**	**33**	**30**	**7**

Partition list into 3 lists: <7 , =7 , >7

3

S_L

7	**7**

S_V

Find the 5th smallest element in the following table:

7	**10**	**3**	**20**	**15**	**16**	**70**	**33**	**30**	**7**

Choose a random element as pivot

7	**10**	**3**	**20**	**15**	**16**	**70**	**33**	**30**	**7**

Partition list into 3 lists: <7 , =7 , >7

3

S_L

7	**7**

S_V

10	**20**	**15**	**16**	**70**	**33**	**30**

S_R

Find the 5th smallest element in the following table:

7	10	3	20	15	16	70	33	30	7

Choose a random element as pivot

7	10	3	20	15	16	70	33	30	7

Partition list into 3 lists: <7 , =7 , >7

3

S_L

7	7

S_V

10	20	15	16	70	33	30

S_R

Find the 2nd smallest element in the following table:

10	20	15	16	70	33	30

Find the 5th smallest element in the following table:

7	10	3	20	15	16	70	33	30	7

Choose a random element as pivot

7	10	3	20	15	16	70	33	30	7

Partition list into 3 lists: <7 , =7 , >7

3

S_L

7	7

S_V

10	20	15	16	70	33	30

S_R

Find the 2nd smallest element in the following table:

10	20	15	16	70	33	30

recurse

The split operation

- Choosing a split value v we divide the set S into three subsets:
  - $S_L=\{x \in S | x<v\}$
  - $S_v=\{x \in S | x=v\}$
  - $S_R=\{x \in S | x>v\}$

After we split, we know how to continue

- We know the size of the three sets: $|S_L|$, $|S_v|$, $|S_R|$.
- If median is in S_v, then we are done.
- Otherwise we continue with either S_L or S_R

Randomized algorithm

- Let S be a set of N different numbers. Suppose we pick a random element of S to be the pivot v. What is the probability that the size of S_L is equal to 10 ?
  - 1/10
  - 9/10
  - 1/N

Randomized algorithm

- Let S be a set of N different numbers. Suppose we pick a random element of S to be the pivot v. What is the probability that the size of S_L is between $\lceil N/4 \rceil$ and $\lceil 3N/4 \rceil$?
  A. About ¼
  B. About ½
  C. About ¾
  D. About 1/N

Lucky splits

- We say that the split of a set S of size N is lucky if
- $\frac{1}{4}N \leq |S_L| \leq (3/4)N$
- Which implies also that $\frac{1}{4}N \leq |S_u| \leq (3/4)N$
- If the split is lucky then the size of the set we operate on decreases by a factor of (3/4)
- In order to reduce the set to all-equal elements we need at most k lucky splits:

$$\left(\frac{3}{4}\right)^k N \leq 1 \;\Rightarrow\; k\log\frac{3}{4} + \log N \leq 0 \;\Rightarrow\; k \geq \frac{\log N}{\log(3/4)}$$

Expected time to first lucky split

- What is the expected number of random splits until we get a lucky split?
  A. 1
  B. 2
  C. 1/2

Expected Running time

$n =$ The number of elements in the input array.

$T(n) =$ The expected running time of the algorithm

$$T(n) \leq n + \frac{1}{2}T(n) + \frac{1}{2}T\left(\frac{3}{4}n\right)$$

Multiply both sides by 2 and rearrange:

$$2T(n) \leq 2n + T(n) + T\left(\frac{3}{4}n\right); \qquad T(n) \leq 2n + T\left(\frac{3}{4}n\right)$$

$$T(n) \le 2n + T\left(\frac{3}{4}n\right)$$

$$T(n) \le 2n + \frac{3}{4}2n + T\left(\left(\frac{3}{4}\right)^2 n\right)$$

$$T(n) \le 2n + \left(\frac{3}{4}\right)2n + \left(\frac{3}{4}\right)^2 2n + T\left(\left(\frac{3}{4}\right)^2 n\right)$$

$$T(n) \le 2n\left(1 + \left(\frac{3}{4}\right) + \left(\frac{3}{4}\right)^2 + \left(\frac{3}{4}\right)^3 + \cdots\right) + T(\le 1)$$

$$T(n) \le 2n \frac{1}{1-(3/4)} = 8n$$

This is an upper bound - the actual constant is smaller.

Before you leave,
please bring me your filled-in
KQS cards

Two styles of random algorithms

- Las Vegas: always gives the correct answer, time to terminate varies from run to run, known bound on expected running time.
- Monte Carlo: Takes a pre-defined time to terminate, gives the correct answer with non-zero probability.

Tip for remembering:
Monte Carlo - outcome not always Correct
Las Vegas - run-time Varies

From Las Vegas to Monte-Carlo

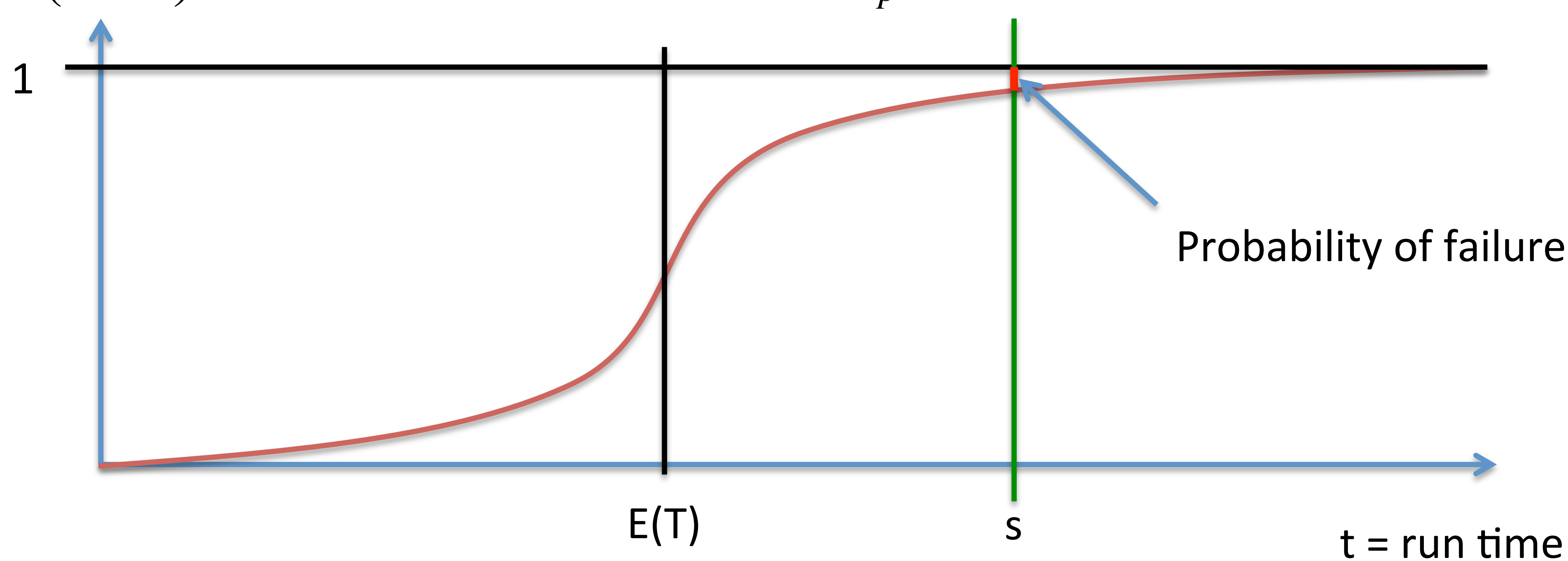
A Las Vegas algorithm always outputs the correct answer but it’s run-time varies, the expected run time is bounded.
We want to transform the algorithm into a Monte-Carlo algorithm that Always takes the same same time to complete but has a small probability p of generating an incorrect answer.
Method: Time out the LV algorithm at time s and output “fail”.

T = the random variable equal to the run-time of the algorithm.

T is non-negative and $E(T)$ is known

Markov inequality: $P(T \geq s) \leq \dfrac{E(T)}{s}$

Prescription: choose time out to be $s \geq \dfrac{E(T)}{p}$



From Monte-Carlo to Las Vegas

A Monte-Carlo algorithm always takes the same same time to complete but has probability p of not generating the correct answer.
We want to transform it into a Las Vegas algorithm, i.e. one that always outputs the correct answer has a varying run time with finite expected value.

Method: run the MC algorithm repeatedly until the answer that it outputs is correct.

Same analysis as the expected number of flips of a biased coin until the first "heads".

$$p = \text{ the probability that MC fails.}$$

$$I = \text{the number of times MC is run until the first correct answer.}$$

$$\Pr(I = i) = p^{i-1}(1-p)$$

$$E(I) = (1-p) + 2p(1-p) + 3p^2(1-p) + \cdots =$$

$$= (1-p)\sum_{i=1}^{\infty} ip^{i-1} = \frac{1-p}{p}\sum_{i=1}^{\infty} ip^{i} = \frac{1-p}{p}\frac{p}{(1-p)^2} = \frac{1}{1-p}$$

Example: What is the expected number of iteration If the MC algorithm Succeeds with probability ¼ ?