

Parallel And Sequential Data Structures and Algorithms

Probability Review and Tail Bounds

Learning Objectives

- Review **basic probability tools** to analyze random processes
 - i.e. sample spaces, events, union bound, random variables, independence, expectation, and linearity of expectation
- Understand **tail-bounds** including Markov's inequality and **high probability bounds**

Motivating Example

The Random Distance Run (RDR)

Organized by SCS and held in April every year

Roll two big fuzzy dice and everyone runs that many laps on the Gesling track.

(note: second is rolled when first person finishes laps from first)



The Random Distance Run (RDR)

Some questions:

- What is the **expected** number of laps?
- What if we took the **product** of the dice?
- Would it matter if the second always rolled the same as the first?
- What if we took the **maximum** of the dice?



Basic Probability

Definitions

Sample Space (Ω): Set of possible outcomes of a “well defined experiment”. We’re doing discrete probability, so Ω is finite or countably infinite.

e.g., $\Omega = \{(1,1),(1,2),\dots,(1,6),(2,1),\dots,(6,5),(6,6)\}$ for the RDR
 $|\Omega| = 36$

Elementary Event: $e \in \Omega$

e.g., $e = (3,5)$ i.e. 3 laps then 5 laps

Event: $A \subseteq \Omega$

e.g., $A = \{(1,3),(3,1),(2,2)\}$ -- dice add to 4
 $B = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}$ -- first dice is 1

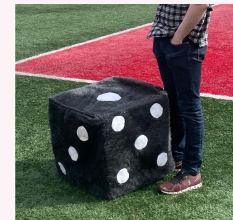
Definitions

Probability Measure (Pr): $Pr : 2^\Omega \rightarrow \mathbb{R}$ (2^Ω is all subsets of Ω)

1. For any event A : $0 \leq Pr(A) \leq 1$,
2. If $A \cap B = \emptyset$ then $Pr(A) + Pr(B) = Pr(A \cup B)$,
3. and $Pr(\Omega) = 1$ (Probabilities add to 1)

e.g. for RDR:

- $Pr(\{(1,1)\}) = 1/36$
- $Pr(\{(1,1), (1,2)\}) = 1/18$
- $Pr(\{(1,1), (1,2), (1,3)\}) = 1/12$
- $Pr(\{(1,1), (1,2), \dots, (1,6), (2,1), \dots, (6,5), (6,6)\}) = 1$



Definitions

Probability Measure (Pr): $Pr : 2^\Omega \rightarrow \mathbb{R}$ (2^Ω is all subsets of Ω)

1. For any event A : $0 \leq Pr(A) \leq 1$,
2. If $A \cap B = \emptyset$ then $Pr(A) + Pr(B) = Pr(A \cup B)$,
3. and $Pr(\Omega) = 1$ (Probabilities add to 1)

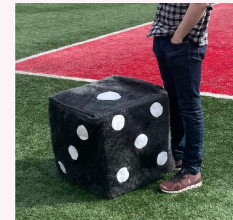
e.g. for RDR check rule 2

Lets say $A = \{(1,1)\}$ and $B = \{(1,2),(1,3)\}$

Then $Pr(A) = 1/36$, $Pr(B) = 1/18$

and $A \cap B = \emptyset$ so

$$Pr(A \cup B) = Pr(\{(1,1),(1,2),(1,3)\}) = 1/36 + 1/18 = 1/12$$



Definitions

Probability Measure (Pr): $Pr : 2^\Omega \rightarrow \mathbb{R}$ (2^Ω is all subsets of Ω)

1. For any event A : $0 \leq Pr(A) \leq 1$,
2. If $A \cap B = \emptyset$ then $Pr(A) + Pr(B) = Pr(A \cup B)$,
3. and $Pr(\Omega) = 1$ (Probabilities add to 1)

For our purposes (with discrete probability) can replace 2. with:

$$Pr(A) = \sum_{x \in A} Pr(x)$$

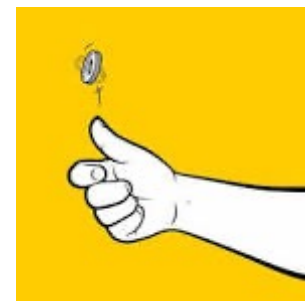
Exercise: prove these are the same.

Another example:

Consider the “experiment” consisting of flipping coins until a head shows up. In this case the outcomes are 1 flip, 2 flips, 3 flips etc. This is an infinite sized set $\Omega = \{1, 2, 3, 4, \dots\}$, but “countable”. Probabilities look like:

$$P(1) = \frac{1}{2}, \quad P(2) = \frac{1}{4}, \quad P(3) = \frac{1}{8}, \dots,$$

They still sum to 1 in the limit.



Conditional Probability

Definition (Conditional Probability): for events A and B :

$$Pr(A | B) = Pr(A \cap B) / Pr(B)$$

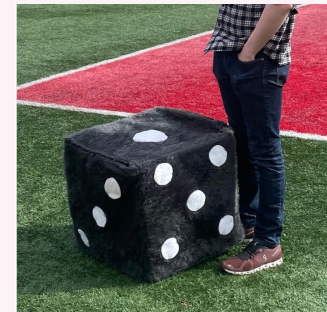
i.e., the probability of event A , given that our outcome is in the set B .

Returning to the Random Distance Run (RDR):

A = first dice is 2

B = dice add to 4

$Pr(A | B) = ?$



$$Pr(A | B) = (1/36) / (3/36) = 1/3$$

Independence

Definition (Independence): Events A and B are **independent** if

$$Pr(A | B) = Pr(A)$$

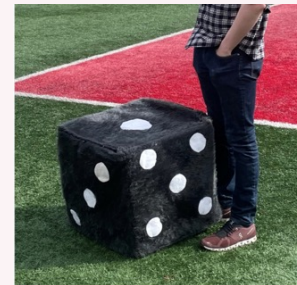
i.e., knowing B gives no information about A

For the RDR, which of the following are independent?

A = first dice is 2

B = dice add to 4

C = second dice is 3



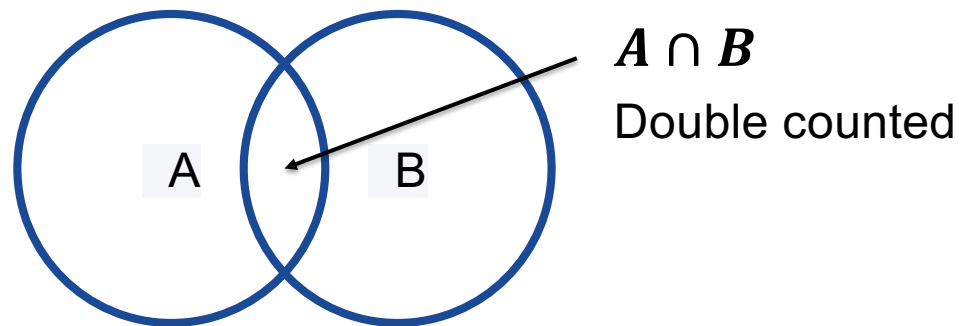
A and C are independent, the others are not.

e.g. $P(A | C) = 1/6 = P(A)$, $P(A | B) = 1/3 \neq P(A)$

Union Bound

Theorem (Union Bound): For any events A and B:
$$Pr(A \cup B) \leq Pr(A) + Pr(B)$$

**Ven Diagram
of the sets A and B:**



$$\begin{aligned} Pr(A \cup B) &= Pr(A) + Pr(B) - P(A \cap B) \\ &\leq Pr(A) + Pr(B) \end{aligned}$$

Union Bound

Theorem (Union Bound): For any events A and B:
$$Pr(A \cup B) \leq Pr(A) + Pr(B)$$

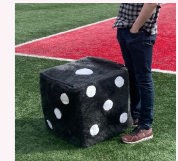
In the RDR:

A = sum to 4

B = first roll is a 2

$$Pr(A) = 1/12, Pr(B) = 1/6 \quad \rightarrow \quad P(A \cup B) \leq 1/4$$

Actual probability of {summing to 4} **or** {first roll is 2} is 8/36.



Random Variable (RV)

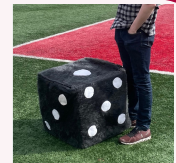
Definition (Random Variable): A function $X: \Omega \rightarrow \mathbb{R}$

In this view **not** a variable, and **not** random.
Apparently, the term is historical from before probability theory.

In the RDR for example:

X = value of first roll

$$X((1,1)) \rightarrow 1, \quad X((1,2)) \rightarrow 1, \quad X((3,5)) \rightarrow 3$$

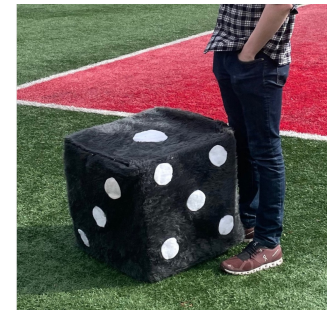


Indicator Random Variable

Definition (Indicator Random Variable): A random variable that is 1 if an event happens and 0 otherwise.

RDR:

$$X = \begin{cases} 1 & \text{dice add to 4} \\ 0 & \text{otherwise} \end{cases}$$



Importantly: $E[X] = \Pr(X = 1)$

Notation

Notation: $(X = a) \equiv \text{the event } \{y \in \Omega \mid X(y) = a\}$

In RDR for example, again assuming RV X = value of first roll
 $(X = 3) \equiv \{(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)\}$

Similarly: $(X > a) \equiv \text{the event } \{y \in \Omega \mid X(y) > a\}$

And:

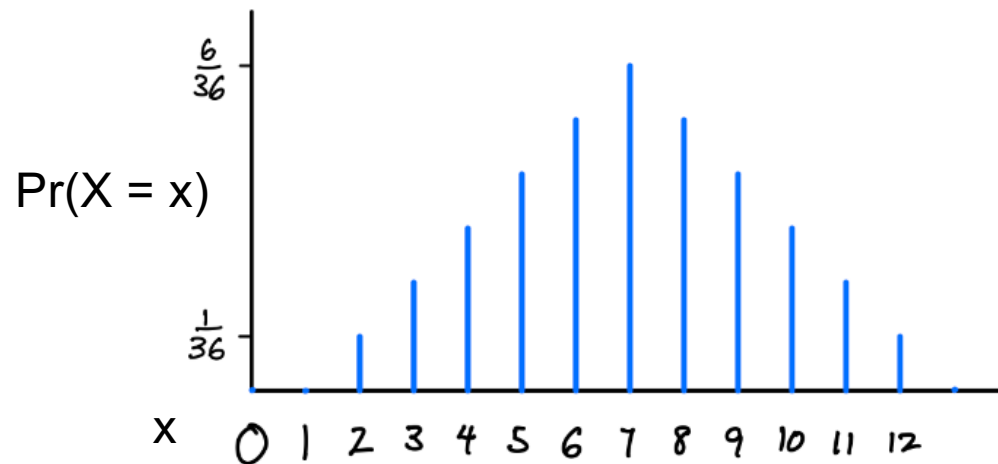
$(X = a, Y = b) \equiv \text{the event } \{z \in \Omega \mid (X(z) = a) \wedge (Y(z) = b)\}$

Probability Density Function of a RV

Probability Density Function of a (discrete) RV:

$$f(x) = \Pr(X = x)$$

e.g. for the RV X = “sum of two dice” we have:



More notational shorthand

Notation (binary operations on random variables):

For random variables X and Y :

$$X + Y \equiv \lambda e. (X(e) + Y(e))$$

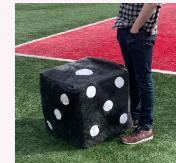
Similarly for other binary operators, e.g. $(-, \times, \max)$

In RDR:

X = value of first roll

Y = value of second roll

$X + Y$ = total laps needed to run



Independence of Random Variables

Definition (RV Independence): Random variables X and Y are independent if

$$Pr(X = a, Y = b) = Pr(X = a)Pr(Y = b) \quad \forall a, b$$

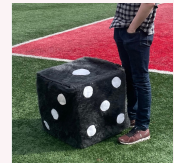
In RDR for example:

X = value of first flip

Y = value of second flip

Z = sum of flips

X and Y are independent, but X and Z are not (exercise)

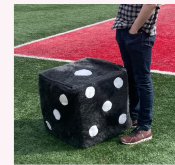


Expectation

Definition (Expectation, $E[X]$): Intuitively: the weighted average of X over all possible outcomes. More formally:

$$\mathbb{E}[X] = \sum_{a \in \Omega} \text{Pr}(a) \cdot X(a) = \sum_x x \cdot \text{Pr}(X = x)$$

In the RDR consider 1 dice with $\Omega = \{1,2,3,4,5,6\}$ and X = value of dice. Then: $\mathbb{E}[X] = \sum_{i \in \{1..6\}} \frac{1}{6} i = \frac{7}{2}$



Coin Flipping Revisited

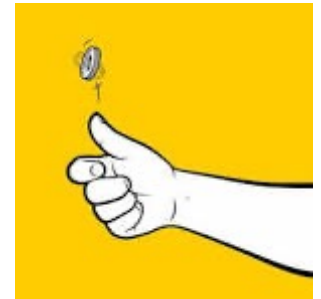
Recall that flipping until a head has
 $\Omega = \{1, 2, 3, 4, \dots\}$, with probabilities

$$P(1) = \frac{1}{2}, \quad P(2) = \frac{1}{4}, \quad P(3) = \frac{1}{8}, \dots,$$

Let a random variable Y be the number of flips until a head is reached. We have:

$$\mathbb{E}[Y] = \sum_{i \in \Omega} \frac{1}{2^i} i = 2$$

Will cover how to solve this later.



Linearity of Expectation

Theorem (Linearity of Expectation): For two random variables X and Y , we have that:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

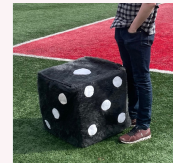
In the RDR, for

X = value of first dice

Y = value of second dice

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = 3\frac{1}{2} + 3\frac{1}{2} = 7$$

i.e. the expected number of laps is 7.



Linearity of Expectation

Theorem (Product of Expectation): For two **independent** random variables X and Y , we have that:

$$E[X \times Y] = E[X] \times E[Y]$$

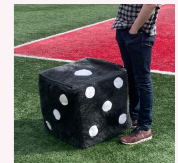
In the RDR, for

X = value of first dice

Y = value of second dice

$$\mathbb{E}[X \times Y] = \mathbb{E}[X] \times \mathbb{E}[Y] = 3\frac{1}{2} \times 3\frac{1}{2} = 12\frac{1}{4}$$

But what if dice are not independent, but always roll the same?



Example : Random Distance Run

Product of two dice, but they always roll the same:

$$\Omega = \{(1,1), (2,2), (3,3), \dots, (6,6)\}$$

$$\mathbb{E}[X \times Y] = \sum_{(e,e) \in \Omega} \frac{1}{6} e^2 = 15 \frac{1}{6}$$

vs. $12 \frac{1}{4}$ when independent.



Example : Random Distance Run

What about if we take the maximum (assume independent)?

No easy way, have to do full sum:

$$\begin{aligned} & \mathbf{E}[\max(X, Y)] \\ &= \sum_{(x,y) \in \Omega} \frac{1}{36} \cdot \max(x, y) = 4 \frac{17}{36} \end{aligned}$$



Summary : Random Distance Run

Sum independent: 7

Sum tied together: 7

Product independent: $12 \frac{1}{4}$

Product tied together: $15 \frac{1}{6}$

Max independent: $4 \frac{17}{36}$

Max tied together: $3 \frac{1}{2}$



Problem with span

For work we can use linearity of expectation, e.g.: for two computations L and R, with work W_L, W_R we have that

parallel(L, R)

yields:

$$\mathbf{E}[W_L + W_R] = \mathbf{E}[W_L] + \mathbf{E}[W_R] \quad (\text{linearity of expectation})$$

But for span S_L, S_R

$$\mathbf{E}[\max(S_L, S_R)] \neq \max(\mathbf{E}(S_L), \mathbf{E}(S_R))$$

Problem with span

Consider 100 students taking a test with no time limit.

1. With probability 99/100 a student takes 1 hour
2. with probability 1/100 a student takes 100 hours

What is the expected time for **a student** to complete?



At most 1.99 hours ($.99 \times 1 + .01 \times 100$)

What is the expected time for the **last student** to complete?

About 63 hours (not trivial to compute).

Tail Bounds

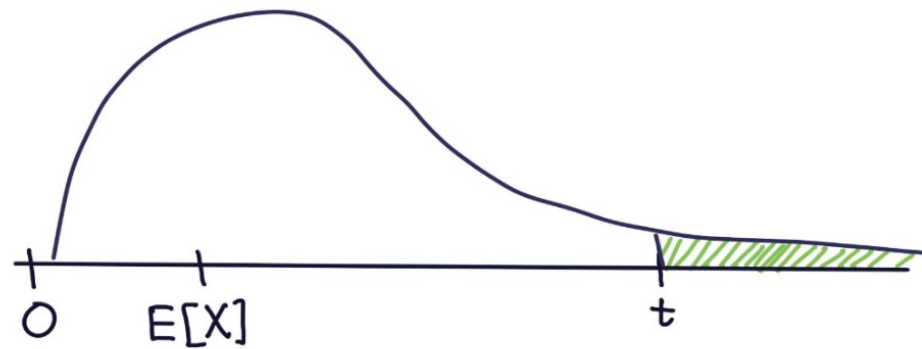
Tail Bounds



Goal: show that $\Pr(X > t)$ is small.

Intuitively: if it is very unlikely (e.g., one in a million) that any student takes more than 2 hours, then it is unlikely the last student among 100 will take more than 2 hours (about 1 in ten thousand)

Tail Bounds



In this class we consider two types of tail bound:

1. **Markov's inequality**
2. **High probability bounds**

Markov's Inequality

Theorem (Markov's Inequality): For a random variable X which is always at least 0, and for $a \geq 1$:

$$\Pr(X \geq aE[X]) \leq \frac{1}{a}$$

In words: it is not very likely to be very much greater than the expectation, and the probability goes down inversely with the factor above the expectation

Markov's Inequality

Theorem (Markov's Inequality): For a random variable X which is always at least 0, and for $a \geq 1$:

$$\Pr(X \geq aE[X]) \leq \frac{1}{a}$$

Intuition: Suppose you have a room full of people for which the average height ($E[H]$) is 5 feet. What percentage can be greater than $2 \times 5 = 10$ feet?

If more than half were greater than 10 feet, then the average would have to be greater than 5 feet, right? Just consider contribution from the 10 feet people.

High Probability Bounds

In these bounds we are going to consider how random variables can grow with a problem size n .

For example, the random variable could be work or span, and we want to place expected or tail bounds on these RVs as a function of n .

High Probability Bounds

Definition: We say that $W(n) \leq f(n)$ **with high probability (w.h.p.)** if there exists a constant n_0 such that for any integer value $k \geq 1$, and any $n \geq n_0$:

$$\Pr(W(n) \leq k \cdot f(n)) \geq 1 - \frac{1}{n^k}$$

We will also write $W(n) \in O(f(n))$ w.h.p. if $W(n) \leq c \cdot f(n)$ for some constant c .

Intuitively it says that as n grows the probability of the costs being within the $f(n)$ upper bound becomes increasingly likely, getting very close to 1 (e.g. consider $n = 10^6$ and $k = 5$).

Max Preserves

Theorem (max preserves w.h.p.): Let $S(n)$ be a non-negative random variable, and let $T(n) = \max(S(n), \dots, S(n))$, where there are n (not necessarily independent) copies of $S(n)$ in the max. If $S(n) \leq f(n)$ w.h.p. then $T(n) \leq 2f(n)$ w.h.p. It follows that if $S(n) \in O(f(n))$ w.h.p. then so is $T(n)$.

Proof:

$$\Pr(T(n) > (k+1)f(n)) \leq n \cdot \Pr(S(n) > (k+1)f(n)) \quad [\text{Union bound}]$$

But the RHS is at most $\frac{n}{n^{k+1}} = \frac{1}{n^k}$ by virtue of $S(n) \leq f(n)$ w.h.p.

Also note that $2k \geq k+1$ so we can write:

$$\Pr(T(n) > 2kf(n)) \leq \Pr(T(n) > (k+1)f(n)) \leq \frac{1}{n^k}$$

which is the definition of $T(n) \leq 2f(n)$ w.h.p.

Max Preserves

Theorem (max preserves w.h.p.): Let $S(n)$ be a non-negative random variable, and let $T(n) = \max(S(n), \dots, S(n))$, where there are n (not necessarily independent) copies of $S(n)$ in the max. If $S(n) \leq f(n)$ w.h.p. then $T(n) \leq 2f(n)$ w.h.p. It follows that if $S(n) \in O(f(n))$ w.h.p. then so is $T(n)$.

Importantly: This allows us to compose the span of randomized algorithms. If we can ensure every of n parallel calls is within some high probability bound (e.g. $O(\log(n))$ with probability $1 - \frac{1}{n^5}$) then the union bound only knocks off 1 from the probability exponent moving us to e.g. $1 - \frac{1}{n^4}$.