Chapter 22

Introduction to Probability Theory

Probability theory is a means of calculating the likelihood of different *events* occurring when conducting some well-defined *experiment*.

Experiments: An experiment might be as simple as flipping a coin and observing whether the event heads or the event tails occurs. It might consist of buying a lottery ticket and observing how much money is gained or lost. It might also consist of executing a randomized algorithm on a given input instance and observing how long it executes and whether it gives the correct output.

Probability of an Event: The probability of event A is a real number $p = \Pr[A] \in [0, 1]$ that measures the fraction of times that the event occurs.

Definition: There are two ways of defining this probability: either by repeating the experiment or by looking into how the experiment works. Either way, it involves counting.

1) Running Many Times: If you repeat the experiment "independently" an infinite number of times N, then the probability of an event p is defined to be the fraction of those times in which the event occurs, that is, pN times out of the N trials.

$$p = \Pr[A] = \lim_{N \to \infty} \frac{\text{The \# of times event } A \text{ occurs in } N \text{ trials}}{N}$$

2) Inner Workings: If we count all possible outcomes r of an experiment where each outcome is equally likely to occur, the probability p of event A can then be defined to be

$$p = \Pr[A] = \frac{\text{The \# of } r \text{ for which event } A \text{ occurs}}{\text{The \# of } r}$$

Underlying Coin Flips: Suppose, for example, that the randomness for the experiment comes from flipping a fair coin a fixed number of times. Then, r might be $\langle heads, tails, heads, heads, \ldots, tails \rangle$.

Random Real in [0, 1]: Computers cannot actually flip coins. Instead, your program can call a system routine which tries to return some thing that is *pseudo random*. A common routine *rand* returns a random real value x between zero and one. You can use this to simulate other random distributions. For example, you can simulate a 6-sided die as follows. If $x \in [0, \frac{1}{6})$ pretend that you rolled a one. If $x \in [\frac{1}{6}, \frac{2}{6})$, pretend you rolled a two and so on. This works because for any $0 \leq a \leq b \leq 1$, $\Pr[x \in [a, b]] = b - a$.

Examples:

Coin Flip: Given this definition, it is easy to see that the probability of the event *heads* when flipping a coin is $p = \frac{1}{2}$.

Dying: To help get perspective on the probability $p = \frac{1}{10{,}000{,}000}$, it is approximately the probability of dying in the next five minutes, because people generally live at most 90 years which is $90 \cdot 365 \cdot 24 \cdot 60/5 \approx 10,000,000$ blocks of 5 minutes and we approximate that you die in a random one of these.

Probability of x successes: Let P_x denote the probability that there are exactly x successes when running n independent experiments each with success probability p.

$$P_x = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!\ (n-x)!} p^x (1-p)^{n-x}.$$

This is because there are $\binom{n}{x}$ ways of "choosing" x of the n experiments to be the ones that will succeed. For each of these ways of choosing, the probability that the x chosen experiments succeed is p^x and the probability that the $n-x$ experiments not chosen fail is $(1-p)^{n-x}$.

Venn Diagrams: A useful way to visualize probabilities is with *Venn Diagrams*. Draw a square with area one. Let each point in it represent one outcome $r = \langle heads, tails, heads, heads, \ldots, tails \rangle$ of the coin flips. For each event, circle those outcomes that lead to the event occurring. The area of the circled region is the probability of the event. For example, Figure 22.1.1 represents the fact that event A occurs with probability $p = \frac{1}{3}$ and fails to occur with probability $1-p = \frac{2}{3}$.

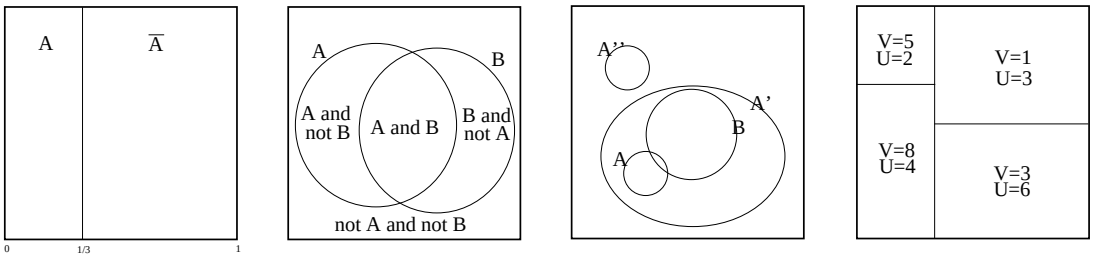


Figure 22.1: The four Venn diagrams. The first shows that the probability of event A is $p = \frac{1}{3}$. The second shows the probability of events A and B happening simultaneously, the probability of only one or the other occurring, and neither occurring. The third demonstrates events A and B being independent, positively dependent, or negatively dependent. The last shows a random variable V with $\Pr[V = 5] = \frac{1}{9}$.

Dependencies Between Events: When you have more than one event, the dependencies between them can be complicated.

Venn Diagrams: Venn diagrams are useful for visualizing these dependencies. Figure 22.1.2 represents the event when both event A and event B both occur simultaneously, when only one or the other occurs, and when neither occurs.

Probability of A given B: The probability of an event A is also related to the extent of our knowledge about whether the event will occur. If all coin flip outcomes r are equally likely, then $\Pr[A]$ tells us the likelihood of event A happening. But suppose now that we knew that event B happened. This narrows the possible coin flip outcomes r to only those for which B occurs. See the B circle in Figure 22.1.2. Given this, the fraction of times that A will be happen is

$$\Pr[A|B] = \frac{\text{The \# of } r \text{ for which both } A \text{ and } B \text{ occur}}{\text{The \# of } r \text{ for which event } B \text{ occurs}} = \frac{\Pr[A \text{ and } B]}{\Pr[B]}$$

Independence and Dependence: Events can be dependent in different ways. Figure 22.1.3 gives examples.

Definition of Independent Events: Events A and B are said to be *independent* if knowing that B occurs, does not give you any information about whether A occurs. For example, when you flip two coins, their outcomes are independent, that is, whether coin 1 is heads or tails does not affect whether coin 2 is heads or tails. The formal definition is

$$\Pr[A|B] = \Pr[A].$$

An equivalent definition is that events A and B are independent if and only if

$$\Pr[A \text{ and } B] = \Pr[A] \cdot \Pr[B].$$

See Exercises 22.0.1 and 22.0.2. Note that this second definition shows the symmetry that if A is independent of B, then B is independent of A.

I drew events A and B in Figure 22.1.3 to be independent events. If the area of the box is one, of A is $\frac{1}{25}$, and of B is $\frac{1}{9}$, then area of the intersection $A \cap B$ must be is $\frac{1}{25} \times \frac{1}{9}$. Given I eyeballed it, I make no promises. If the same two circles were moved so they overlapped ever so slightly more, then the event A and B would be positively dependent, while if they were moved to overlap ever so slightly less, then they would be negatively dependent.

Positively Dependent: Events A' and B are said to be *positively dependent* if they are more likely to occur together, that is $\Pr[A'|B] > \Pr[A']$ and $\Pr[A' \text{ and } B] > \Pr[A'] \cdot \Pr[B]$. (Note that in Figure 22.1.3, $\Pr[A'|B] = 1$.) Even if event A' occurs if and only iff event B occurs, we do not really know why this happens. This may occur because B "causes" A to happen, because A "causes" B to happen, or because some event C "causes" both A and B to happen. A butterfly flapping its wings in Africa and a storm in Toronto are likely independent events, but they say that in this interconnected chaotic world, these events may be dependent.

Negatively Dependent: Events A'' and B are said to be *negatively dependent* if they are less likely to occur together, that is $\Pr[A''|B] < \Pr[A'']$ and $\Pr[A'' \text{ and } B] > \Pr[A''] \cdot \Pr[B]$. (Note that in Figure 22.1.3, $\Pr[A''|B] = 0$.)

Random Variables: Some experiments result in a value, like your win-

nings at gambling or the running time of a randomized algorithm. The resulting value V is referred to as a *random variable*, as it takes on different values with different probabilities.

Examples:

Venn Diagram: In Figure 22.1.4, $\Pr[V = 5] = \frac{1}{9}$ and $\Pr[V = 1] = \frac{1}{3}$.

Number of Heads: If you flip a coin n times, the number of times that you get a head is a random variable. If you flip it 4 times, V can take on values between 0 and 4. $\Pr[V = 2] = \frac{3}{8}$ and $\Pr[V = 4] = \frac{1}{16}$

Indicator Variables: An *indicator variable* I_A is a random variable which is 1 when the event A being indicated occurs and zero when it does not.

Running Time: The running time T of a randomized algorithm is a random variable.

Expected Value: The *expected value* of a random variable is not the value that you expect, but is the average value if you were to repeat it many times.

Definition: The following are three equivalent definitions.

Average: Suppose again that the randomness comes from flipping a fair coin a fixed number of times and let V_r denote the value of V when the outcomes of the coin flips is r. Each r is equally likely to occur. The expected value of V is its average value.

$$\text{Exp}[V] = \frac{\Sigma_r \, V_r}{\text{The \# of different } r} = \sum_r \Pr[r] V_r$$

Value: A more standard definition considers separately each value v that V might take on.

$$\text{Exp}[V] = \sum_{\text{values } v} \Pr[V = v] \cdot v$$

Disjoint Events: Sometimes it is easier to partition the universe of possible outcomes into a set of events of your choosing. As in the "Average" definition of expected value, an event could be that the coins came up as r. As in the "Value" definition of expected value, an event could be that random variable V takes on the value v. Or you can come up with your own set of events that make will make your calculations as easy as possible.

$$\text{Exp}[V] = \sum_{\text{disjoint events } A} \Pr[A] \cdot [\text{value of } V \text{ during event } A]$$

.

Examples:

Coin Flip: If you get $V = 1$ for a head and $V = -1$ for a tail, then the expected amount is $\text{Exp}[V] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot (-1) = 0$.

Venn Diagram: In Figure 22.1.4, the expected value of V is $\text{Exp}[V] = \Sigma_v \Pr[V = v] \cdot v = \frac{1}{9} \cdot 5 + \frac{1}{3} \cdot 1 + \frac{2}{9} \cdot 8 + \frac{1}{3} \cdot 3 = 3\frac{2}{3}$.

Lotteries: If you pay \$5 for a lottery ticket and with probability $p = \frac{1}{10{,}000{,}000}$ you win \$25,000,000, then your expected winnings are $(1 - \frac{1}{10{,}000{,}000}) \cdot 0 + \frac{1}{10{,}000{,}000} \cdot 25{,}000{,}000 = \$2.50$. But you paid \$5. Hence, you expect to lose half your money. This is surprising, because I expect you will lose all of your money.

Expected Happiness: Money is not everything though. What is your expected gain in happiness? I claim having \$5 given your current level of wealth adds more to your happiness than

having \$5 when you already have \$25,000,000. This proves that happiness does not increase linearly with money. In fact, I would guess it is more logarithmic because no matter how much you have, if the amount you have doubles, your happiness increases by more or less a fixed amount. So let's guess that buying a roti with your \$5 would bring you one unit of happiness and winning \$25,000,000 would bring you 1,000 units of happiness. You say more? Okay, 100,000 units. Then your expected happiness gained by buying a ticket is $(1 - \frac{1}{10{,}000{,}000}) \cdot (-1) + \frac{1}{10{,}000{,}000} \cdot 100,000 \approx (-1) + 0.001 \approx -1$, i.e. you lose.

Expected Number: If you flip a coin n times, the expected number of times that you get a head is $\frac{n}{2}$.

Indicator Variables: The expected value of an indicator variable I_A equals the probability of the event A, that is $\text{Exp}[I_A] = \Pr[A] \cdot 1 + \Pr[\text{not } A] \cdot 0 = \Pr[A]$.

Linearity of Sum of Expectation: A very useful fact is that the expectation of the sum is equal to the sum of the expectations. Let $V_1, V_2, V_3, \ldots, V_n$ be n random variables, which may or may not be dependent in complicated ways. If you form a new random variable denoted V' whose value on every outcome of the coins is the sum of the V_i, then

$$\text{Exp}\left[V'\right] = \text{Exp}\left[\sum_i V_i\right] = \sum_i \text{Exp}\left[V_i\right].$$

Venn Diagram: In Figure 22.1.4,

$\text{Exp}[V] = \Sigma_v \Pr[V = v] \cdot v = \frac{1}{9} \cdot 5 + \frac{1}{3} \cdot 1 + \frac{2}{9} \cdot 8 + \frac{1}{3} \cdot 3 = 3\frac{2}{3}$.

$\text{Exp}[U] = \Sigma_u \Pr[U = u] \cdot u = \frac{1}{9} \cdot 2 + \frac{1}{3} \cdot 3 + \frac{2}{9} \cdot 4 + \frac{1}{3} \cdot 6 = 4\frac{1}{9}$.

$\text{Exp}[(V+U)] = \Sigma_w \Pr[(V+U) = w] \cdot w = \frac{1}{9} \cdot 7 + \frac{1}{3} \cdot 4 + \frac{2}{9} \cdot 12 + \frac{1}{3} \cdot 9 = 7\frac{7}{9}$.

We can check that $\text{Exp}[(V+U)] = 7\frac{7}{9} = 3\frac{2}{3} + 4\frac{1}{9} = \text{Exp}[V] + \text{Exp}[U]$.

Proof: The proof that the expectation of the sum is equal to the sum of the expectations is as follows. The formal definition of the expectation is $\text{Exp}[U+V] = \Sigma_w \Pr[U+V=w] \cdot w$, however, it is not clear what to do with this. It is better to break the universe of possibilities into finer events. For every tuple $\langle u, v \rangle$, consider the event that $U = u$ and $V = v$. Note that when this event occurs, we know that the random variable $[U+V]$ takes on the value $u + v$. This gives that

$$\begin{aligned}\text{Exp}[U+V] &= \sum_{\substack{\text{disjoint events } A}} \Pr[A] \cdot [\text{value of } [U+V] \text{ during} \\ &= \sum_{\langle u,v \rangle} \Pr[U = u \text{ and } V = v] \cdot (u+v)\end{aligned}$$

The distributive and the commutative laws gives that $[p \cdot (u+v)] + [p' \cdot (u'+v')] = [pu + pv] + [p'u' + p'v'] = [pu + p'u'] + [pv + p'v']$. Such rearranging gives

$$\begin{aligned}\text{Exp}[U+V] &= \left[\sum_{\langle u,v \rangle} \Pr[U = u \text{ and } V = v] \cdot u\right] + \\ &\quad \left[\sum_{\langle u,v \rangle} \Pr[U = u \text{ and } V = v] \cdot v\right]\end{aligned}$$

Think of a matrix of values indexed by u and v. The sum of the entries can be obtained by summing them up. It can also be obtained by summing each row and then summing these sums or by summing each column and then summing these sums.

$$\begin{aligned}\text{Exp}[U+V] &= \sum_u \left[\sum_v \Pr[U = u \text{ and } V = v] \cdot u\right] + \\ &\quad \sum_v \left[\sum_u \Pr[U = u \text{ and } V = v] \cdot v\right]\end{aligned}$$

We now use the reverse of the distributive law, $pu + p'u = (p + p')u$.

$$\begin{aligned}\text{Exp}[U+V] \;=\; & \sum_u \left[\sum_v \Pr\left[U = u \text{ and } V = v\right]\right] \cdot u + \\ & \sum_v \left[\sum_u \Pr\left[U = u \text{ and } V = v\right] \cdot v\right]\end{aligned}$$

Fix some value u. What is $\Sigma_v \Pr[U = u \text{ and } V = v]$? If you think of the Venn diagram, $\Pr[U = u]$ is the area of the union of all the areas in which $U = u$. In some of those areas, $V = v$ and in some of them $V = v'$. It follows that $\Sigma_v \Pr[U = u \text{ and } V = v] = \Pr[U = u]$. Hence,

$$\text{Exp}[U+V] \;=\; \sum_u \Pr[U = u] \cdot u + \sum_v \Pr[V = v] \cdot v$$

but by the definition of expected values this gives

$$\text{Exp}[U+V] \;=\; \text{Exp}[U] + \text{Exp}[V]$$

Expected Number of Successes: If you have n trials where each trial has success with probability p, the expected number of successes is pn. This is true even if the success of each trial dependent in complicated ways on each other.

Proof: A simple proof is as follows.

$$\begin{aligned}\text{Exp}\left[\text{Numb of successes}\right] &= \text{Exp}\left[\sum_i I_i\right] = \sum_i \text{Exp}\left[I_i\right] \\ &= \sum_i \left[p \cdot 1 + (1-p) \cdot 0\right] = pn\end{aligned}$$

Expected Time Till Success: If I flip a fair coin until I get a head, I may have to flip it only once or a million times, but the expected number of times I have to flip it is two. If I roll a dice until I get a six, the expected number of times I have to

roll it is six. More generally, suppose an experiment succeeds with probability p. Suppose I repeat it independently until it succeeds. Let the random variable T be the number of times that it is repeated. A not too surprising but useful lemma is that $\text{Exp}[T] = \frac{1}{p}$.

Proof 1: For T to equal the value t, it requires that the experiment fails the first $t-1$ time and then succeeds the t^{th} time. The probability of this is $\Pr[T=t] = (1-p)^{t-1}p$. This gives that $\text{Exp}[T] = \Sigma_{t=1}^{\infty} \Pr[T=t]t = \Sigma_{t=1}^{\infty}(1-p)^{t-1}p \cdot t$. This is a really hard sum to evaluate (Ask if you want me do it for you). It does, however, add up to $\frac{1}{p}$ as we want.

Proof 2: This proof is hard too. Skip it if you like. For each $t \geq 0$, let I_t be an indicator variable which is 1 if you must repeat the experiment more than t times.

- What is $\Pr[I_t=1]$?
  - Answer: You will need to repeat the experiment more than t times only if it failed the first t times. The probability of this is $\Pr[I_t] = (1-p)^t$.
- What is $\text{Exp}[I_t]$?
  - Answer: $\text{Exp}[I_t] = \Pr[I_t] = (1-p)^t$.
- What is T in terms of the I_t?
  - Answer: $T = \Sigma_{t\geq 0} I_t$ is the total number of experiments tried.
- What is $\text{Exp}[T]$? Hint: For $0 \leq q < 1$, $\Sigma_{t\geq 0} q^t = \frac{1}{1-q}$.
  - Answer: $\text{Exp}[T] = \text{Exp}[\Sigma_{t\geq 0} I_t] = \Sigma_{t\geq 0} \text{Exp}[I_t] = \Sigma_{t\geq 0}(1-p)^t = \frac{1}{p}$.

Expectation of Product: The same thing is true for the product of random variable if the random variables are independent and is

not necessarily true if they are dependent.

Proof when Independent: We prove as follows that if $V_1, V_2, V_3, \ldots, V_n$ are independent random variables, then

$$\text{Exp}\left[V'\right] \;=\; \text{Exp}\left[\Pi_i V_i\right] = \Pi_i \text{Exp}\left[V_i\right].$$

The proof begins the way it did for the sum of expectations.

$$\text{Exp}[U \times V] \;=\; \sum_w \Pr\left[U \times V = w\right] \cdot w \;=\; \sum_u \sum_v \Pr\left[U = u \text{ and } V\right.$$

Because the events are independent we have that $\Pr\left[U = u \text{ and } V = v\right] \;=\; \Pr\left[U = u\right] \;\times\; \Pr\left[V = v\right]$. Then commutativity gives $(p \times p') \cdot (u \times v) = (p \cdot u) \times (p' \cdot v)$.

$$\text{Exp}[U \times V] \;=\; \sum_u \sum_v \left[\Pr\left[U = u\right] \cdot u\right] \times \left[\Pr\left[V = v\right] \cdot v\right]$$

We now use the distributed law that $pq + pq' + p'q + p'q' = (p + p') \times (q + q')$.

$$\text{Exp}[U \times V] \;=\; \left[\sum_u \Pr\left[U = u\right] \cdot u\right] \times \left[\sum_v \Pr\left[V = v\right] \cdot v\right] = \text{Exp}\left[U\right.$$

Proof when Not Independent: We prove as follows that if the random variables are dependent than the previous result is not necessarily true.
Suppose that $V_1 = V_2 = 0$ with probability $\frac{1}{2}$ and $V_1 = V_2 = 2$ with probability $\frac{1}{2}$. Then $\text{Exp}[V_1] = \text{Exp}[V_2] = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 2 = 1$. $\text{Exp}[V_1 \cdot V_2] = \frac{1}{2} \cdot (0 \cdot 0) + \frac{1}{2} \cdot (2 \cdot 2) = 2$. This is different than $\text{Exp}[V_1] \cdot \text{Exp}[V_2]$.

Random Walks: Consider a line (side walk with squares) with a wall at $i = 0$ and a wall at $i = n$.

Completely Drunk: Each time step, the drunk man is standing at some square i and with probability $\frac{1}{2}$ stumbles one square forward

and with probability $\frac{1}{2}$ stumbles one square backwards. When $i = 0$, he only goes forward. What is the expected number of time steps starting at $i = 0$ until the man to first gets to $i = n$. Guess. Is it $2n$, n^2, 2^n or something else?

Proof: Let t_i denote the expected number of time steps starting at square i until the man to first gets to square $i+1$. We can write a recurrence relation. With probability $\frac{1}{2}$, he goes forward and it takes him only one step. However, with probability $\frac{1}{2}$, he goes backwards and that one step takes him to square $i - 1$. From here, the expected number of time steps until he first returns to square i is t_{i-1}. From here, the expected number of time steps until he first gets to square $i+1$ is t_i. Because the expectation of the sum is the sum of the expectations, we get the following

$$t_i = \tfrac{1}{2}[1] + \tfrac{1}{2}[1 + t_{i-1} + t_i]$$

$$\tfrac{1}{2}t_i = 1 + \tfrac{1}{2}t_{i-1}$$

$$t_i = 2 + t_{i-1} = 2 + 2 + t_{i-2} = 2j + t_{i-j} = 2i + t_0 = 2i + 1$$

.

The expected number of time steps starting at $i = 0$ until the man to first gets to $i = n$ is the expected number of time steps until he first gets to $i = 1$ plus the the expected number until he first gets from there to $i = 2$ and so on, which is

$$\sum_{i=0}^{n-1} t_i = \sum_{i=0}^{n-1} 2i + 1 = n^2 + \Theta(1).$$

Smelling Home: Now suppose that he stumbles forward with probability $\frac{1}{2} + \epsilon$ and backwards with with probability $\frac{1}{2} - \epsilon$. We want to know how much better this guy does.

Proof: Let W_t^ϵ denote the random variable giving the index i of where the man is at time t. Similarly, let W_t denote the same but when the probabilities are half and half. Then $W_t^\epsilon - W_t$ is the random variable denoting how far ahead the smelling man is from the drunk man. If the randomness of the two men are independent, then it is hard to compare their locations. Instead, let us *couple* their probabilities. Divide the unit line into three pieces of lengths $\frac{1}{2} - \epsilon$, ϵ, and $\frac{1}{2}$. Each step, we throw one dart. If it lands in the first interval, we call this B and both men move back one square. If it lands in the second, we call this ϵ and the drunk man move back one square and the smelling man moves forward one square. If it lands in the third interval, we call this F and both men move forward one square. Note that the distance $W_t^\epsilon - W_t$ increases by two in the second case and stays fixed in the other two.

$$W_t = \#F - (\#\epsilon + \#B)$$

$$W_t^\epsilon = (\#F + \#\epsilon) + \#B$$

(Note if $\epsilon = \frac{1}{2}$, then $W_t^\epsilon = (\#F + \#\epsilon) + \#B) = (\#F + \#\epsilon) = t$, because $\Pr[B] = 0$.)

$$W_t^\epsilon - W_t = 2\#\epsilon$$

$$W_t^\epsilon = W_t + 2\#\epsilon$$

$$\mathrm{Exp}[W_t^\epsilon] = \mathrm{Exp}[W_t] + 2\epsilon t$$

We can now state that the expected time until the smelling man reaches $i = n$ from $i = 0$ is less than $\min(\frac{n}{2\epsilon}, n^2 + \Theta(1))$. If $\epsilon >> \frac{1}{n}$, then in $\frac{n}{2\epsilon}$ time, we can't expect the drunk man to have gotten very far, but we can expect the smelling man to be

n steps in front of the him and hence past the $i = n$ line. On the other hand, if $\epsilon << \frac{1}{n}$, then in $n^2 + \Theta(1))$ time, we can't expect the smelling man to have gotten very far ahead of the drunk man, but we can expect the drunk man to have reached the $i = n$ line and so so will have the smelling man.

Plotting Probability: Other useful ways to visualize random variables are using the following three functions.

Value from Point: A Venn graph like Figure 22.1.4 labels each point in the unit square with a real number. You can imagine throwing a dart at the unit square uniformly at random (meaning each point in the square is equally likely to get hit. The value of the random variable V will be the real value labeling the unit square at that point. Using the unit square has the advantage that you can dawn it nicely as done in Figure 22.1.4. However, instead of a unit square, you could just as easily use the unit line. We will use function $\widehat{V} : [0, 1] \Rightarrow \mathcal{R}$ to label each real value point x in the unit interval $[0, 1]$ with a real number. You can imagine throwing a dart at the unit interval uniformly at randomly obtaining some real value x. The value of the random variable V will be the real value $\widehat{V}(x)$ labeling the unit interval at that point x. In general, $\widehat{V}$ can be an arbitrary function, but for are purposes here, we might as well assume that the values V are sorted so that $\widehat{V}(x)$ is a non-decreasing function.

Figure 22.1.4: For example, the random variable V in Figure 22.1.4. has $\widehat{V}(x) = 1$ for $x \in [0, \frac{1}{3}]$, $\widehat{V}(x) = 3$ for $x \in [\frac{1}{3}, \frac{2}{3}]$, $\widehat{V}(x) = 5$ for $x \in [\frac{2}{3}, \frac{7}{9}]$, and finally $\widehat{V}(x) = 8$ for $x \in [\frac{7}{9}, 1]$.

Real in [0,6 :] As as second example, let $\widehat{V}(x) = 6x$ and then V is a random variable that uniformly takes on a random real

value from 0 to 6. Note, that because V can take on any real from 0 to 6, the probability it takes on any particular value like 2 is effectively zero. On the other hand, $\Pr[V \leq v]$ is $\frac{2}{6} = \frac{1}{3}$.

$\Pr[V \leq v]$: The second function $P_{(\leq)} : \mathcal{R} \Rightarrow [0,1]$ used to describe a random variable V is defined to be

$$P_{(\leq)}(v) = \Pr[V \leq v].$$

Increasing: Note that $P_{(\leq)}(-\infty) = \Pr[V \leq -\infty] = 0$. Then $P_{(\leq)}(v)$ increases with v until $P_{(\leq)}(\infty) = \Pr[V \leq \infty] = 1$.

Figure 22.1.4: For example, the random variable V in Figure 22.1.4. has $P_{(\leq)}(v) = 0$ for $v \in [0,1)$, $P_{(\leq)}(v) = \frac{1}{3}$ for $v \in [1,3)$, $P_{(\leq)}(v) = \frac{2}{3}$ for $v \in [3,5)$, $P_{(\leq)}(v) = \frac{7}{9}$ for $v \in [5,8)$, and $P_{(\leq)}(v) = 1$ for $v \in [8,\infty)$.

Real in [0,6 :] When V is a random variable that uniformly takes on a random real value from 0 to 6, then for $v \in [0,6]$, $P_{(\leq)}(v) = \Pr[V \leq v] = \frac{v}{6}$.

$P_{(\leq)}$ Inverse of $\widehat{V}$: Suppose that the previously mentioned function $\widehat{V}$ is strictly increasing. Hence, if $\widehat{V}(x) = v$, then $\widehat{V}(x') \leq v$ for all $x' \in [0,x]$ and $\widehat{V}(x') > v$ for all $x' \in (x,1]$. This gives that $P_{(\leq)}(v) = \Pr[V \leq v] = x$ and hence that $P_{(\leq)}(\widehat{V}(x)) = x$, i.e. $\widehat{V}$ and $P_{(\leq)}$ are inverses of each other. If $\widehat{V}$ is non decreasing, but could take on the same value for a while, then it is a little trickier, but one can show that $\widehat{V}(P_{(\leq)}(v)) = v$.

$\Pr[V = v]$: The third function $P_{=}$ used to describe a random variable V is defined to express $\Pr[V = v]$.

Discrete V: If the random variable V takes on discrete values $v_1, v_2, \ldots, v_r$, then a *histogram* has a place in the X axis for each of the possible values $v_1, v_2, \ldots, v_r$, and above v_i is a bar

of width one and height (and area) $\Pr[V = v_i]$. Denote the resulting curve by $P_{(=)}$. Note that the "area" of under this curve is one because $\Sigma_i \Pr[V = v_i]$ must be one.

Figure 22.1.4: For example, the random variable V in Figure 22.1.4. has $P_=(1) = \frac{1}{3}$, $P_=(3) = \frac{1}{3}$, $P_=(5) = \frac{1}{9}$, and $P_=(8) = \frac{2}{9}$.

Continuous V: If the random variable V takes a range of real values, then doing a histogram is more complicated because then $\Pr[V = v]$ is effectively zero.

Infinitesimals: What we will do instead is break the range of values v into intervals each of width δv, where δv is your favorite some infinitesimal value. Then instead of considering $\Pr[V = v]$, we consider $\Pr[V \in [v, v+\delta v]]$. Though this probability is still an infinitesimal, we can still imagine this being bigger than zero.

Histogram: We will now build a histogram, just as we did in the discrete case. It has a place in the X axis for each of the v intervals. Above v is a bar of width δv, area $\Pr[V \in [v, v+\delta v]]$, and height $\frac{\Pr[V \in [v, v+\delta v]]}{\delta v}$. Denote the resulting curve by $P_{(=)}$.

Real in [0,6 :] When V is a random variable that informally takes on a random real value from 0 to 6, then $P_{(=)}(v) = \frac{\Pr[V \in [v, v+\delta v]]}{\delta v} = \frac{\delta v/6}{\delta v} = \frac{1}{6}$. This curve is constant ($P_{(=)}(v) = \frac{1}{6}$), which is the case for uniform distributions.

$\Pr[V \in [v_1, v_2]]$: From this curve we can read off any probability, because

$$\Pr[V \in [v_1, v_2]] = \sum_{intervals\ v \in [v_1, v_2]} \Pr[V \in [v, v+\delta v]$$

$$= \sum_{intervals\ v \in [v_1, v_2]} P_{(=)}(v)\delta v = \int_{v \in [v_1, v_2]} P_{(=)}(v)\delta v,$$

which is the area under the curve from v_1 to v_2. Therefore, the area under the entire curve is $\Pr[V \in [-\infty, \infty]] = 1$.

$P_{(\leq)}$: Note this gives a relationship between this last two functions for expressing the random variable V.

$$P_{(\leq)}(v) = \Pr[V \leq v] = \int_{v \in [-\infty, v]} P_{(=)}(v)\delta v,$$

which is the area under the curve to the left of value v. Conversely $P_{(=)}$ is the derivative (slope) of $P_{(\leq)}(v)$, because

$$\frac{\delta P_{(\leq)}(v)}{\delta} = \frac{P_{(\leq)}(v+\delta v) - P_{(\leq)}(v)}{\delta} = \frac{\Pr[V \in [v, v+\delta v]]}{\delta v} = I$$

Markov's Tail Inequality: If V is a random variable that only takes on non-negative values and v is any fixed value, then

$$\Pr[V \geq v] \leq \frac{\text{Exp}[V]}{v}$$

Proof: Let V be a random variable that only takes on non-negative values and v is any fixed value. Let X be the random variable which equals v if $V \geq v$ and 0 otherwise. $\text{Exp}[V] \geq \text{Exp}[X] = v \cdot \Pr[V \geq v]$. Rearranging give that $\Pr[V \geq v] \leq \frac{\text{Exp}[V]}{v}$.

Silly Example: In Figure 22.1.4, $0.2222 = \frac{2}{9} = \Pr[V \geq 8] \leq \frac{\text{Exp}[V]}{v} = \frac{3\ 2/3}{8} = 0.4583$.

Uses: Often in practice, we can compute one of $\Pr[V \geq v]$ or $\text{Exp}[V]$ but not both. Markov's Inequality can be used to approximate the other.

Standard Deviation: $\text{Exp}[V]$ gives the expected or the average value of the random variable V. However, we might also want to know how likely or how much the actual value of V deviates far from this expectation, namely $|V - \text{Exp}[V]|$. We could compute the expected deviation, that is $\text{Exp}[|V - \text{Exp}[V]|]$, however, the absolute values make the

computations cumbersome. Hence, we compute the expected value of the square of the deviation, namely

$$\begin{aligned}\text{Variance}[V] &= \text{Exp}\left[(V-\text{Exp}[V])^2\right] \\ &= \sum_v \Pr[V=v]\cdot(v-\text{Exp}[V])^2.\end{aligned}$$

The square acts like it is taking the absolute value because both negative and positive values become positive. Another effect of squaring the deviation is that large deviations like $V-\text{Exp}[V]=100$ when squared become even more significant. The next thing that we do to take the square root of this expected value, because if V is in units of, say, meters, then so is $(V-\text{Exp}[V])$, but $(V-\text{Exp}[V])^2$ and $\text{Exp}\left[(V-\text{Exp}[V])^2\right]$ would be meters squared. By taking the square root of this, the units become meters again. We call this the *standard deviation* of the random variable V.

$$\text{StandardDiviation}[V] = \sqrt{\text{Exp}[(V-\text{Exp}[V])^2]}$$

Example:

Balanced: Suppose that $V=2$ with probability $\frac{1}{2}$ and $V=8$ with probability $\frac{1}{2}$. Its expected value is $\text{Exp}[V]=\frac{1}{2}\cdot 2+\frac{1}{2}\cdot 8=5$, its variance is $\text{Var}[V]=\Sigma_v \Pr[V=v]\cdot(v-\text{Exp}[V])^2=\frac{1}{2}\cdot(2-5)^2+\frac{1}{2}\cdot(8-5)^2=\frac{1}{2}\cdot(-3)^2+\frac{1}{2}\cdot(+3)^2=9$, and its standard deviation is $SD[V]=\sqrt{\text{Var}[V]}=3$. This makes sense because we expect V to deviate by 3 from its expected value 5.

Venn Diagram: In Figure 22.1.4, the expected value of V is $\text{Exp}[V]=3\frac{2}{3}$, its variance is $\text{Var}[V]=\Sigma_v \Pr[V=v]\cdot(v-\text{Exp}[V])^2=\frac{1}{9}\cdot(5-3\frac{2}{3})^2+\frac{1}{3}\cdot(1-3\frac{2}{3})^2+\frac{2}{9}\cdot(8-3\frac{2}{3})^2+\frac{1}{3}\cdot(3-3\frac{2}{3})^2=6\frac{8}{9}$ and its standard deviation is $SD[V]=\sqrt{\text{Var}[V]}=2.624...$

Another Expression for Variance:

$$\text{Variance}[V] = \text{Exp}\left[V^2\right]-\text{Exp}[V]^2$$

Proof: $\text{Variance}[V] = \text{Exp}\left[(V - \text{Exp}[V])^2\right] =$
$\text{Exp}\left[V^2 - 2V\text{Exp}[V] + \text{Exp}[V]^2\right] = \text{Exp}\left[V^2\right] -$
$2\text{Exp}[V]\text{Exp}[V] + \text{Exp}[V]^2 = \text{Exp}\left[V^2\right] - \text{Exp}[V]^2$

Linearity of Variance: If V and U are two independent random variables, then

$$\text{Variance}[V + U] = \text{Variance}[V] + \text{Variance}[U].$$

Proof: $\text{Variance}[V + U] = \text{Exp}\left[([V + U] - \text{Exp}[V + U])^2\right]$
(the expectation of the sum is the sum of the expectation, then rearrange)
$= \text{Exp}\left[([(V - \text{Exp}[V])] + [(U - \text{Exp}[U])])^2\right]$
$= \text{Exp}\left[(V - \text{Exp}[V])^2 + 2(V - \text{Exp}[V]) \cdot (U - \text{Exp}[U]) + (U -\right.$
$= \text{Exp}\left[(V - \text{Exp}[V])^2\right] + \text{Exp}\left[(U - \text{Exp}[U])^2\right]$
$+2\text{Exp}\left[(V - \text{Exp}[V]) \cdot (U - \text{Exp}[U])\right]$
(for independent random variables the expectation of the product is the product of the expectation.)
$= \text{Variance}[V] + \text{Variance}[U] + 2\text{Exp}\left[V - \text{Exp}[V]\right] \cdot$
$\text{Exp}\left[U - \text{Exp}[U]\right]$
$= \text{Variance}[V] + \text{Variance}[U] + 2[\text{Exp}[V] - \text{Exp}[V]] \cdot [\text{Exp}[U] -$
$\text{Exp}[U]]$
$= \text{Variance}[V] + \text{Variance}[U] + 2[0] \cdot [0].$

Trials: Let V be the random variable indicating the number of successes when you have n independent trials where each trial has success with probability $p \leq \frac{1}{2}$. You expect to get pn successes. We will show that the variance is close to pn giving that the standard deviation is $\sqrt{pn}$.

Proof: Let I_i be the *indicator variable* which is 1 when the i^{th} of the trials succeeds and 0 otherwise. Hence, the number of successes is $V = \Sigma_i I_i$. $\text{Exp}[I_i] = p \cdot 1 + (1 - p) \cdot 0 = p$. $\text{Variance}[I_i] =$

$\text{Exp}\left[(I_i - \text{Exp}[I_i])^2\right] = p \cdot (1-p)^2 + (1-p) \cdot (0-p)^2 = p(1-p)$.
$\text{Variance}[V] = \Sigma_i \text{Variance}[I_i] = p(1-p)n$. But we assume p is small, so this is close to p.

Chebyshev's Tail Inequality: If V is a random variable (taking on positive or negative values) and h is any fixed value, then

$$\Pr[|V - \text{Exp}[V]| \geq h] \leq \frac{SD[V]^2}{h^2}$$

Proof: Let V be a random variable and h is any fixed value. Let $Y = (V - \text{Exp}[V])^2$ be a random variable. By definition, $\text{Exp}[Y] = SD[V]^2$. Hence, by Markov's inequality $\Pr[|V - \text{Exp}[V]| \geq h] = \Pr\left[Y \geq h^2\right] \leq \frac{\text{Exp}[Y]}{h^2} = \frac{SD[V]^2}{h^2}$.

Silly Example: In Figure 22.1.4, $0.2222 = \frac{2}{9} = \Pr[V \geq 8] \leq \Pr[|V - \text{Exp}[V]| \geq 8 - 3\frac{2}{3}] \leq \frac{SD[V]^2}{h^2} = \frac{(2.624..)^2}{(8-3\frac{2}{3})^2} = 0.3666$.

Uses: Knowing only the expectation $\text{Exp}[V]$ one can use Markov's Inequality to approximate and event. Knowing the standard deviation as well, one can improve this approximation.

Chernoff's Tail Inequalities: Let V be the random variable indicating the number of successes when you have n independent trials where each trial has success with probability $p \leq \frac{1}{2}$. You expect to get pn successes. You won't likely get exactly pn successes, but you are likely to get within a few standard deviations of this. Here the standard deviation is $\sqrt{pn}$. The probability of deviating farther from this is exponentially small.

Deviating by h:

$$\Pr[V \leq pn - h] \leq e^{-h^2/(2pn)}.$$

$$\Pr[V \geq pn + h] \leq e^{-h^2/(2(pn+h))}.$$

Deviating by a Constant Factor: For example, the probability of getting a constant factor fewer, that is, $h = \epsilon pn$, is exponentially small.

$$\Pr[V \leq pn - \epsilon pn] \leq e^{-\epsilon^2 pn/2} = e^{-\Theta(n)}.$$

Deviating by a c Standard Deviations: My favorite way of expressing it is as follows. The standard deviation is $\sqrt{pn}$. The probability of getting c standard deviations too few, that is, $h = c\sqrt{pn}$, is at most $e^{-c^2/2}$.

$$\Pr[V \leq pn - c\sqrt{pn}] \leq e^{-c^2/2}.$$

For example, if you flip a fair coin 20,000 times, the probability of getting fewer than $pn - 6\sqrt{np} = \frac{1}{2}20,000 - 600 = 9,400$ heads is at most $e^{-c^2/2} = e^{-6^2/2} \approx 10^{-8}$. Similarly, if you flip it a large n number times, than the fraction of heads is very likely at most $\frac{n/2+6\sqrt{n/2}}{n} \approx \frac{1}{2}$.

Proof Sketch: We start by shifting our random variable V to V' that its expectation is zero. Let I_i be the shifted *indicator variable* which is $1-p$ when the i^{th} of the trials succeeds and $-p$ otherwise. Hence, when the number of successes is $V \geq pn + h$, we have that $V' = \Sigma_i I_i \geq (pn + h)(1 - p) + [n - (pn + h)](-p) = [pn(1 - p) - (1 - p)np] + [h(1 - p) + hp] = h$. Let t be some value to be optimized later. Remember that when random variables are independent, the expectation of their product (Π_i) is the product of their expectation.

$$\Pr[V \geq h] = \Pr[e^{tV} \geq e^{th}] \leq \frac{\mathrm{Exp}[e^{tV}]}{e^{th}} = \frac{\mathrm{Exp}[e^{\Sigma_i tI_i}]}{e^{th}} = \frac{\mathrm{Exp}[\Pi_i e^{tI}}{e^{th}}$$

$$= \frac{\Pi_i \text{Exp}[e^{tI_i}]}{e^{th}} = \frac{\Pi_i [p \cdot e^{t\cdot(1-p)} + (1-p) \cdot e^{t\cdot(-p)}]}{e^{th}} = \frac{[pe^t + (1-p)]^n}{e^{th+tpn}}$$

Because this is true for every choices of t, one just has to set t to minimize this probability.

Probability of Succeeding at Least Once: Suppose that that your experiment, say the running of an algorithm, succeeds with at least probability p. Suppose that you are able to repeat the experiment independently N times and that you only need to succeed at least one of these times to succeed over all. Finally, suppose that you want to succeed overall with probability $1 - \epsilon$ for some small $\epsilon > 0$. Then it is sufficient to repeat the experiment $N = \frac{1}{p} \ln\left(\frac{1}{\epsilon}\right)$ times. The probability that you fail each of these times is at most

$$\Pr[AlwaysFail] \leq (1-p)^N \leq e^{-pN} = e^{-\ln\left(\frac{1}{\epsilon}\right)} = \epsilon.$$

For example, if $p = \frac{1}{n^2}$ and $\epsilon = 10^{-9}$ (one in a billion), then $N = \frac{1}{p} \ln\left(\frac{1}{\epsilon}\right) = n^2 \ln\left(10^9\right) \leq 21n^2$. See Exercise **??**.

Probability of a Bad Event: Suppose that there is a list of bad things that might happen. Suppose that you can prove that the probability that the i^{th} one happens is at most p_i. It follows that

$$\Pr[\text{At least one bad thing happens}] \leq \sum_i p_i$$

Proof: Suppose the probability that the i^{th} bad thing happens is at most p_i. The worst case is when these bad events are disjoint so that two never occur simultaneously. Imagine a Venn diagram with disjoint circles of area p_i. In this case, the probability that one happens is exactly $\Sigma_i \, p_i$. More formally, Pr[At least one bad thing happens] =

$$\frac{\text{The \# of } r \text{ for which at least one bad thing happens}}{\text{The \# of } r}$$
$$\leq \Sigma_i \frac{\text{The \# of } r \text{ for which the } i^{th} \text{ bad thing occurs}}{\text{The \# of } r} = \Sigma_i \, p_i.$$

Some Useful Approximations:

$1 - p \leq e^{-p}$: This is useful bound that we have seen already. It is very close to equality when p is close to zero, i.e. $1 - 0 = 1 = e^{-0}$. Here are some other similar inequalities.

- $1 - p + \frac{p^2}{2} \geq e^{-p}$
- $1 + p \leq e^p$ and for $p \in [0, 1]$, $1 + p + p^2 \geq e^p$ and $1 + p \geq e^{p - \frac{p^2}{3}}$.
- $(1 - p)^n = 1 - np + \Theta(p^2)$.
- $1 + p \leq \frac{1}{1-p}$ and very close when p is small.

$n! \approx \left(\frac{n}{e}\right)^n$: This is a fairly close approximation of $n!$ which is the number of ways of arranging n objects. Stirling's approximation, which is even closer, is $n! = \sqrt{e\pi n} \cdot \left(\frac{n}{e}\right)^n e^w$ where $\frac{1}{12(n+.5)} \leq w \leq \frac{1}{12n}$.

$\left(\frac{n}{a}\right)^a \leq \binom{n}{a} \leq \left(\frac{en}{a}\right)^a$: This is a fairly close approximation of $\binom{n}{a} = \frac{n!}{a!(n-a)!}$ which is the number of subsets of size a of n objects. Another approximation, which is even closer is $\binom{n}{rn} \approx 2^{Entropy(r) \cdot n}$, where $Entropy(r) = r \log_2 \frac{1}{r} + (1-r) \log_2 \frac{1}{1-r}$.

Proofs:

$1 + p \leq e^p$ and $1 - p \leq e^{-p}$: These come from two formal definitions of the base of natural logarithms 2.718.., i.e. $\lim_{N \to \infty}(1 + \frac{1}{N})^N = e$ and $\lim_{N \to \infty}(1 - \frac{1}{N})^N = e^{-1}$. Also if you plot the two functions $1 + x$ and e^x using the fact that the derivative of both $1 + x$ and e^x at zero is 1, you can see that $1 + x \leq e^x$ and similarly $1 - x \leq e^{-x}$. These approximations are very close when $x \in o(1)$.

$1+p+p^2 \geq e^p$ and $1-p+\frac{p^2}{2} \geq e^{-p}$: The Taylor expansion of a function $f(x)$ at point x_0 is a close approximation of $f(x)$ for x close to x_0. It is defined to be $f(x_0+p) \approx f(x_0) + f'(x_0)p + \frac{1}{2!}f''(x_0)p^2 + \frac{1}{3!}f'''(x_0)p^3 + \ldots$. Hence, $e^p \approx e^0 + e^0 p + \frac{1}{2!}e^0 p^2 + \frac{1}{3!}e^0 p^3 + \ldots = 1 + p + \frac{1}{2!}p^2 + \frac{1}{3!}p^3 + \ldots \leq 1 + p + p^2$ when $p \leq 1$. Replacing p with $-p$ gives $e^{-p} \approx 1 + (-p) + \frac{1}{2!}(-p)^2 + \frac{1}{3!}(-p)^3 + \ldots \leq 1 - p + \frac{p^2}{2}$.

$(1-p)^n = 1 - np + \Theta(p^2)$: You know that $(1-p)^2 = 1-2p+p^2$ and $(1-p)^3 = 1-3p+3p^2-p^3$. More generally, $(a+b)^n = \Sigma\, i = 0^n \binom{n}{i} a^{n-i} b^i$ and hence $(1-p)^n = \Sigma\, i = 0^n \binom{n}{i}(-p)^i = 1 - np + \frac{n^2}{2}p^2 - \Theta(p^3)$.

$1 + p \leq \frac{1}{1-p}$: $\frac{1}{1-p} = 1 + p + p^2 + p^3 + \ldots$. The p^2 become small when p is small.

$n! \approx \left(\frac{n}{e}\right)^n$: $\ln(n!) = \ln(1 \cdot 2 \cdot 3 \cdot \ldots \cdot n) = \ln(1) + \ln(2) + \ln(3) + \ldots + \ln(n) = \Sigma_{i=1}^{n} \ln(i) \approx \int_{i=1}^{n} \ln(i) = n\ln(n) - n$. Hence, $n! = e^{\ln(n!)} = e^{n\ln(n)-n} = \left[e^{\ln(n)}\right]^n \cdot e^{-n} = \frac{n^n}{e^n}$.

$\left(\frac{n}{a}\right)^a \leq \binom{n}{a} \leq \left(\frac{en}{a}\right)^a$: $\binom{n}{a} = \frac{n!}{a!(n-a)!} = \frac{n(n-1)(n-2)\ldots(n-a+1)}{a(a-1)(a-2)\ldots 1} = \frac{n}{a} \cdot \frac{n-1}{a-1} \cdot \frac{n-2}{a-2} \cdot \ldots \cdot \frac{n-a+1}{1} \geq \left(\frac{n}{a}\right)^a$. $\binom{n}{a} = \frac{n!}{a!(n-a)!} = \frac{n(n-1)(n-2)\ldots(n-a+1)}{a!} \leq \frac{n^a}{a!} \approx \frac{n^a}{(a/e)^a} = \left(\frac{en}{a}\right)^a$.

Exercise 22.0.1 *(See solution in Section ??) Prove that the two definitions of independent events are equivalent, namely* $\Pr[A|B] = \Pr[A]$ *and* $\Pr[A \text{ and } B] = \Pr[A] \cdot \Pr[B]$.

Exercise 22.0.2 *(See solution in Section ??) When A and B are independent, compute* $\Pr[A \text{ and not } B]$ *and* $\Pr[\text{not } A \text{ and not } B]$ *in terms of* $\Pr[A]$ *and* $\Pr[B]$.

Exercise 22.0.3 *(See solution in Section ??) Prove that these three definitions of* $\text{Exp}[V]$ *are equivalent.*

Exercise 22.0.4 *(See solution in Section ??) Suppose* V *is a random variable that only takes on values in the range* $[0, M]$*. Use Markov's inequality to prove the following.*

- $\Pr[V < v] \geq 1 - \frac{\text{Exp}[V]}{v}$
- $\Pr[V \leq v] \leq \frac{M - \text{Exp}[V]}{M - v}$
- $\Pr[V > v] \geq \frac{\text{Exp}[V] - v}{M - v}$

Exercise Solutions

22.0.1 Clearly, the statement $\Pr[A|B] = \frac{\Pr[A \text{ and } B]}{\Pr[B]} = \Pr[A]$ is true if and only the statement $\Pr[A \text{ and } B] = \Pr[A] \cdot \Pr[B]$ is true.

22.0.2 $\Pr[A \text{ and not } B] = \Pr[A] - \Pr[A \text{ and } B] = \Pr[A] - \Pr[A] \cdot \Pr[B] = \Pr[A] \cdot (1 - \Pr[B])$.
$\Pr[\text{not } A \text{ and not } B] = \Pr[\text{not } B] - \Pr[A \text{ and not } B] = (1 - \Pr[B]) - \Pr[A] \cdot (1 - \Pr[B]) = (1 - \Pr[A]) \cdot (1 - \Pr[B])$

22.0.3 $\text{Exp}[V] = \Sigma_{[\text{disjoint events } A]} \Pr[A] \cdot [\text{value of } V \text{ during event } A]$
$= \Sigma_v \Sigma_{[\text{disjoint events } A \text{ for which } V = v]} \Pr[A] \cdot v$
$= \Sigma_v \Pr[V = v] \cdot v.$

Obtaining the coin flips r is like an event A with $\Pr[A] = \frac{1}{\text{The \# of } r}$.
Hence,
$\text{Exp}[V] = \Sigma_{\text{disjoint events } A} \Pr[A] \cdot [\text{value of } V \text{ during event } A] = \Sigma_r \frac{1}{\text{The \# of } r} \cdot V_r.$

22.0.4 • Markov's inequality is $\Pr[V \geq v] \leq \frac{\text{Exp}[V]}{v}$.

- The events $V \geq v$ and $V < v$ are complementary events. Hence, $\Pr[V < v] = 1 - \Pr[V \geq v]$, which by Markov's inequality $\geq 1 - \frac{\text{Exp}[V]}{v}$.
- Let $W = M - V$ be how far V is from it's maximum value. Note that W is a random variable that only takes on non-negative values. Similarly, let $w = M - v$. Then $\Pr[V \leq v] = \Pr[M - V \geq M - v] = \Pr[W \geq w]$, which by Markov's inequality $\leq \frac{\text{Exp}[W]}{w} = \frac{M - \text{Exp}[V]}{M - v}$.
- The events $V \leq v$ and $V > v$ are complementary events. Hence, $\Pr[V > v] = 1 - \Pr[V \leq v]$, which by previous $\geq 1 - \frac{M - \text{Exp}[V]}{M - v} = \frac{\text{Exp}[V] - v}{M - v}$.

Chapter 23

Randomized Algorithms

For some computational problems, allowing the algorithm to flip coins (i.e. use a random number generator) makes for a simpler, faster, makes for a simpler, faster, easier to analyze algorithm. The following are the three main reasons.

Hiding the Worst Cases from the Adversary: The "running time" of a randomized algorithms is analyzed in a different way than that of a deterministic algorithm. At times, this way is more fair and more in line with how the algorithm actually performs in practice. Suppose, for example, that a deterministic algorithm quickly gives the correct answer on most input instances, yet is very slow or gives the wrong answer on a few instances. Its running time and its correctness is generally measured to be that on these worst case instances. A randomized algorithm might also sometimes be very slow or gives the wrong answer. See Quick Sort Section **??**. However, we accept this, as long as on every input instance, the probability of doing so (over the choice of random coins) is small.

Probabilistic Tools: The field of probabilistic analysis has many useful techniques and lemmas that can make the analysis of the algorithm simple and elegant.

Solution has a Random Structure: When the solution that we are

attempting to construct has a random structure, a good way to construct it is to simply flip coins to decide how to built each part. Sometimes we are then able to prove that with high probability the solution obtained this way has better properties than any solution we know how to construct deterministically. Moreover, if we can prove that the solution constructed randomly has extremely good properties with some very small but non-zero probability, for example $prob = 10^{-100}$, then this proves the existence of such a solution even though we have no reasonably quick way of finding one. Another interesting situation is when the randomly constructed solution very likely has the desired properties, for example with probability 0.999999, however, there is no quick way of testing whether what we have produced has the desired properties.

This chapter considers these ideas further.

23.1 Using Randomness to Hide The Worst Cases

The standard way of measuring the running time and correctness of a deterministic algorithm is based on the worst case input instance chosen by some nasty adversary who has studied the algorithm in detail. This is not fair if the algorithm does very well on all but a small number of very strange and unlikely input instances. On the other hand, knowing that the algorithm works well on most instances is not always satisfactory, because for some applications it is just those the hard instances that you want to solve. In such cases, it might be more comforting to use a randomized algorithm that guarantees that on every input instance, the correct answer will be obtained quickly with high probability.

A randomized algorithm is able to flip coins as it proceeds to decide what actions to take next. Equivalently, a randomized algorithm

A can be thought of as a set of deterministic algorithms $A_1, A_2, A_3, \ldots$ where A_r is what algorithm A does when the outcome of the coin flips is $r = \langle heads, tails, heads, heads, \ldots, tails \rangle$. Each such deterministic algorithm A_r will have a small set of worst case input instances on which it either gives the wrong answer or runs too slow. The idea is that these algorithms $A_1, A_2, A_3, \ldots$ have different sets of worst case instances. This randomized algorithm is good if for each input instance, the fraction of the deterministic algorithms $A_1, A_2, A_3, \ldots$ for which it is not a worst case instance is at least p. Then when one of these A_r is chosen randomly, it solves this instance quickly with probability at least p.

I sometimes find it useful to consider the analysis of randomized algorithms as a game between an algorithm designer and an adversary who tries to construct input instance which will be bad for the algorithm. In the game, it is not always fair for the adversarial input chooser to know the algorithm first, because then it can choose the instance that is worst case for this algorithm. Similarly, it is not always fair for the algorithm designer to know the input instance first or even which instances are likely, because then it can design the algorithm to work well on these. The way we analyze the running time of randomized algorithms compromises between these two. In this game, the algorithm designer without knowing the input instance must first fix what his algorithm will do given the outcome of the coins. Knowing this, but not knowing the outcomes of the coins, the instance chooser chooses the worst case instance. We then flip coins, run the algorithm, and see how well it does.

Three Models: The following are formal definitions of three models.

Deterministic Worst Case: In a worst case analysis, a deterministic algorithm A for a computational problem P must always give the correct answer quickly.

$$\forall I,\ [A(I) = P(I) \text{ and } Time(A, I) \leq T_{upper}(|I|)]$$

Las Vegas: The algorithm is said to be *Las Vegas* if the algorithm is always guaranteed to give the correct answer, but the running time of the algorithm depends on the outcomes of the random coin flips. The goal is to prove that on every input instance, the expected running time is small.

$$\forall I,\ [\forall r,\ A_r(I) = P(I) \text{ and } \text{Exp}_r\left[Time(A_r, I)\right] \leq T_{upper}(|I|)]$$

Monte Carlo: The algorithm is said to be *Monte Carlo* if the algorithm is guaranteed to stop quickly, but it can sometimes, depending on the outcomes of the random coin flips, give the wrong answer. The goal is to prove that on every input, the probability of it giving the wrong answer is small.

$$\forall I,\ [Pr_r\left[A_r(I) \neq P(I)\right] \leq p_{fails} \text{ and } \forall r,\ Time(A_r, I) \leq T_{upper}(|I|)]$$

The following examples demonstrate these ideas.

Quick Sort: Recall the quick sort algorithm from Section **??**. The algorithm chooses a pivot element and partitions the list of numbers to be sorted into those that are smaller than the pivot and those that are larger than it. Then it recurses on each of these two parts. The running time varies from $\Theta(n \log n)$ to $\Theta(n^2)$ depending on the choices of pivots.

Deterministic Worst Case: A reasonable choice for the pivot is to always use the element that happens to be located in the middle of array to be sorted. For all practical purposes, this would likely work great. It would work exceptionally well when the list is already sorted. However, there are some strange inputs cooked ups for the sole purpose of being nasty to this particular implementation of the algorithm on which the algorithm runs in $\Theta(n^2)$ time. The adversary will provide such an input giving a worst case time complexity of $\Theta(n^2)$.

Las Vegas: In practice, what is often done is to choose the pivot element randomly from the input elements. This makes it irrelevant which order the adversary puts the elements in the input instance. The expected computation time is $\Theta(nlogn)$.

The Game Show Problem: The input I to the game show problem specifies which of N doors has prizes behind them. At least half the doors are promised to have prizes. An algorithm A is able to look behind the doors in any order that it likes, but nothing else. It solves the problem correctly when it finds a prize. The running time is the number of doors opened.

Deterministic Worst Case: Any deterministic algorithm fixes the order that it looks behind the doors. Knowing this order, the adversary places no prizes behind the first $\frac{N}{2}$ doors looked behind.

Las Vegas: In contrast, a random algorithm will look behind doors in random order. It does not matter where the adversary puts the prizes, the probability that one is not found after t doors is $\frac{1}{2^t}$ and the expected time until a prize is found is $\text{Exp}[T] = \Sigma_t \Pr[T = t] \cdot t = 2$.

Monte Carlo: If the promise is that either at least half the doors have prizes or none of them do and if the algorithm stops after 10 empty doors and claims that there are no prizes, then this algorithm is always fast, but gives the wrong answer with probability $\frac{1}{2^{10}}$.

Randomized Primality Testing: An integer x is said to be *composite* if it has factors other than one and itself. Otherwise, it is said to be *prime*. For example, $6 = 2 \times 3$ is composite and $2, 3, 5, 7, 11, 13, 17, \ldots$ are prime. See Appendix **??** Example 2 for explanations of why it takes $2^{\Theta(n)}$ time to factor an n bit number.[1] Here we give an easy randomized algorithm by Rabin-Miller for this problem.

Fermat's Little Theorem: Don't worry about the math, but Fermat's Little Theorem says that if x is prime, then for every $a \in [1, x-1]$, it is the case that $a^{x-1} \equiv_{(mod\ x)} 1$.

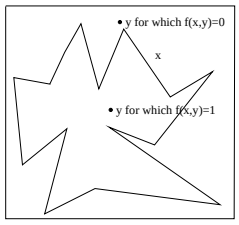
If we want to test if x is prime, then we can pick random a's in the interval and see if the equality holds. If the equality does not hold for a value of a, then x is composite. If the equality does hold for many values of a, then we can say that x is probably prime, or a what we call a *pseudo prime.*

The Game Show Problem: Finding an a for which $a^{x-1} \not\equiv_{(mod\ x)} 1$ is like finding a prize behind door a. See Exercise 23.1.1.

Randomized Counting: In many applications, one wants to count the number of occurrences of something. This problem can often be expressed follows. Given the input instance x, count the number of y for which $f(x, y) = 1$. It is likely very difficult to deter-

[1] A major break through in 2002 Agrawal etal. was to find a polynomial time deterministic algorithm for determining whether an n bit number is prime.

mine the exact number. However, a good way to approximate this number is to randomly choose some large number of values y. For each, test whether $f(x, y) = 1$. Then the fraction of y for which $f(x, y) = 1$ can be approximated by [the number you found]/[the number you tried]. The number of y for which $f(x, y) = 1$ can be approximated by [the fraction you found]×[the total number of y]. For example, suppose you had some strange shape and you wanted to find its area. Then x would specify the shape, y would specify some point within a surrounding box, and $f(x, y) = 1$ if the point is within the shape. Then the number of y for which $f(x, y) = 1$ gives you the area of your shape.



Exercise 23.1.1 *Given an integer x, suppose that you have one door for each $a \in [1, x-1]$. We will say that there is a prize behind this door if $a^{x-1} \not\equiv_{(mod\ x)} 1$. Fermat's Little Theorem says that if x is pseudo prime, then none of the doors have prizes behind them and if it is composite then at least half the doors have prizes. The algorithm attempts to determine which is the case by opening t randomly chosen doors for some integer t.*

1. *If the algorithm finds a prize, what do you know about the integer? If it does not find a prize, what do you know?*
2. *If the algorithm must always give the correct answer, how many doors need to be opened in terms of the number of digits n in the instance x.*
3. *If t doors are open and the input instance x is a pseudo prime, what is the probability that the algorithm gives the correct answer?*

If the instance is composite, what is this probability?

Exercise 23.1.2 *Section* **??** *designed an iterative algorithm for separating n VLSI chips into those that are "good" and those that are "bad" by test two chips at a time and learning either that they are the same or that they are different. To help, at least half of the chips are promised to be good. Now design much easier a randomized algorithm for this problem. Here are some hints.*

- *Randomly select one of the chips. What is the probability that the chip is good?*
- *How can you learn whether or not the selected chip is good?*
- *If it is good, how can you easily partition the chips into good and bad chips.*
- *If the chip is not good, what should your algorithm do?*
- *When should the algorithm stop?*
- *What is the expected running time of this algorithm?*

23.1.1 Sorry no answer

23.1.2 Sorry no answer

23.2 Locker Room Problem

Problem: There are n players, each with a locker and a driver's license. The coach randomly permutes the licenses and puts one in each locker. The players can agree on a strategy. Each player independently goes into the locker room and can look in half the lockers. We say that he succeed if he finds his own license. We say that they succeed if each player succeeds

to finds his own license. They are not allowed to change the room set up or communicate in any way. The probability that a given player succeeds is $\frac{1}{2}$. If things were completely independent then the probability that all succeeds would be $\frac{1}{2^n}$. Is it possible for the players to have a strategy in which they all succeed with a significantly higher probability, say 0.3?

Strategy: Each player starts by looking in his own locker. If he finds Bob's license, he looks in Bob's locker. If in Bob's locker he finds John's license, he looks in John's locker next. This continues until either he finds his locker or has looked in half the lockers.

Permutation Graph: Put a directed edge from i to j if the locker i contains license j. Having out–degree one and in–degree one, this graph contains a collection of cycles.

Success: Player i starts at node i, i.e. his own locker, and follows the edges of this graph. He succeeds when he finds his own driver's license, i.e. when the cycle he is following points back to node i, i.e. he arrives back at node i. Hence, he succeeds when the cycle that he is in contains at most half the nodes. They all succeed if the permutation graph contains no cycles of length greater than half.

Probability of a k Cycle: Let $k \in [\frac{n}{2}+1, n]$. We will show that the probability that a random permutation graph contains a k cycle is $\frac{1}{k}$.

The number of permutation graphs is $n!$ because it can be described by a permutation. There are n choices for a neighbor for node 1 and then $n-1$ choices for a neighbor for node 2, because they can't have the same neighbor, and so on.

Now let us count the number permutations with a cycle of length k. Choose a start node i_1. There are n ways. Choose its neighbor i_2. There are $n-1$ ways, because we don't want to allow node i_1. Choose i_2's neighbor i_3. There are $n-2$ ways, because we don't want to allow nodes

i_1 or i_2. Continue until you choose i_{k-1}'s neighbor i_k. There are $n-(k-1)$ ways. Because we want a cycle of length k, we know that i_k's neighbor is node i_1. Then there $(n-k)!$ ways to arranging the remaining $n-k$ players. The total number of ways is $n!$. However, we over counted by a factor of k because it does not matter which of the k nodes in the k cycle that we started with. Note that we would have over counted further if there was a second cycle of length k in the remaining $n-k$ nodes, but this is not possible because $n-k < k$. Hence, the total number of permutation graphs with a cycle of length k is $\frac{n!}{k}$. The fact that the probability is $\frac{1}{k}$ follows.

Probability of a Large Cycle: There can't be two cycles of more than half the nodes. Hence, the event of there being a k cycle is disjoint for the different $k \in [\frac{n}{2}+1, n]$. Hence the probability of there being a more than half cycle is $\Sigma_{k=\frac{n}{2}+1}^{n} \frac{1}{k} = \Sigma_{k=1}^{n} \frac{1}{k} - \Sigma_{k=1}^{\frac{n}{2}} \frac{1}{k} \approx \ln(n) - \ln(\frac{n}{2}) = \ln(2)$. Hence, the probability of no such large cycle and hence of success is $1-\ln(2) > 0.3$.

23.3 Solutions of Optimization Problems with a Random Structure

Optimization problems are looking for the best solution for an instance. Sometimes good solutions have a random structure. In such cases, a good way to construct it is to simply flip coins to decide how to built each part. We give two examples. The first one, *Max Cut*, being NP-complete, likely requires exponential time to find the best solution. However, in $\mathcal{O}(n)$ time, we can find a solution which is likely to be at least half as good as optimal. The second example, *expander graphs* is even more extreme. Though there are deterministic algorithms for constructing graphs with fairly good expansion properties, a random graph almost for sure has much better expansion properties (with probability $p \geq 0.999999$). A

complication, however, is that there is no polynomial time algorithm which tests whether this randomly constructed graph has the desired properties. Pushing the limits further, it can be proved that the same random graph has extremely good properties with some very small but non-zero probability (eg. $p \geq 10^{-100}$). Though we have no quick way to construct such a graph, this does proves that such a graph exists.

The Max Cut Problem: The input to the Max Cut problem is an undirected graph. The output is a partition of the nodes into two sets U and V so that the number of edges that cross over from one side to the other is as large as possible. This problem is NP-complete and hence, the best known algorithm for finding an optimal solution requires $2^{\Theta(n)}$ time. The following randomized algorithm runs in time $\Theta(n)$ and is expected to obtain a solution for which half the edges cross over. This algorithm is incredibly simple. It simply flips a coin for each node to decide whether to put it into U or into V. Each edge will cross over with probability $\frac{1}{2}$. Hence, the expected number of edges to cross over is $\frac{|E|}{2}$. The optimal solution cannot have more than all the edges cross over, so the randomized algorithm is expected to perform at least half as well as the optimal solution can do.

Expander Graphs: An n node degree d graph is said to be an *Expander Graph* if moving from a set of its nodes across its edges expands us out to an even larger set of nodes. More formally, for $0 < \alpha < 1$ and $1 < \beta < d$, a graph $G = \langle V, E \rangle$ is an $\langle \alpha, \beta \rangle$-expander if for every subset $S \subseteq V$ of its nodes, if $|S| \leq \alpha n$ then $|N(S)| \geq \beta |S|$. Here $N(S)$ is the neighborhood of S, that is all nodes with an edge from some node in S.

Non-Overlapping Sets of d Neighbors: Because each node $v \in V$ has d neighbors $N(v)$, a set S has $d|S|$ edges leaving

these nodes. However, if these sets $N(v)$ of neighbors overlap a lot, then the total number of neighbors $N(S) = \cup_{v \in S} N(v)$ of S might be very small. We can't expect $N(S)$ to be bigger than $d|S|$ but we do want it to have size at least $\beta|S|$ where $1 < \beta < d$. If S is too big, we can't expect it to expand further. Hence, we only require this expansion property for sets S of size at most αn. Because we do expect sets of size αn to expand to a neighborhood of size $\beta\alpha n$, we do require that $\alpha\beta < 1$.

Connected with Short Paths: If $\alpha\beta > \frac{1}{2}$, then every pair of nodes in G is connected with a path of length at most $\frac{2\log(n/2)}{\log\beta}$.

Proof: Consider two nodes u and v. The node u has d neighbors, $N(u)$. These neighbors $N(u)$ must have at least $\beta|N(u)| = \beta d$ neighbors $N(N(u))$. These neighbors $N(N(u))$ must have at least $\beta^2 d$ neighbors. It follows that there are at least $\beta^{i-1}d$ nodes with distance i from u. The last time we are allow to do this expands the neighbor set of size $|S| = \alpha n$ to $|N(S)| \geq \beta|S| = \beta\alpha n$. By the requirement that $\alpha\beta > \frac{1}{2}$, this new neighbor set has size greater than $\frac{n}{2}$ nodes. The distance of these nodes from u is at most $i = \log_\beta \frac{n}{2}$. This set might not contain v. However, starting from v there is another set of more than half the nodes that are distance $i = \log_\beta \frac{n}{2}$ from v. These two sets must overlap at some node w. Hence, there is a path from u to w to v of length at most $\frac{2\log(n/2)}{\log\beta}$.

Uses: Expander graphs are very useful both in practice and for proving theorems.

Fault Tolerant Networks: As we have seen every pair of nodes in an expander graph are connected. This is still true if a large number of nodes or edges fail. Hence, this is a good

pattern for wiring a communications network.

Pseudo Random Generators: Taking a short random walk in an expander graph quickly gets you to a random node. This is useful for generating long random looking strings from a short seed string.

Concentrating and Recycling Random Bits: If we have a source that has some randomness in it (say n coin tosses with an unknown probability and with unknown dependencies between the coins), we can use expander graphs to produce a string of m bits appearing to be the result of m fair and independent coins.

Error Correcting Codes: Expander graphs are also useful in designing ways of encoding a message into a longer code so that if any reasonable fraction of the longer code is corrupted, the original message can still be recovered. The the faulty bits are connected by short paths to correct bits.

If $\alpha\beta < 1$, then Expander Graphs Exists: We will now prove that for any constants α and β for which $\alpha\beta < 1$ there exists an $\langle\alpha, \beta\rangle$-expander graph with n nodes and degree d for some sufficiently big constant d. For example, if $\alpha = \frac{1}{2}$, $\beta = \frac{3}{2}$, then $d = 5$ is sufficient. To make the analysis easier, we will consider directed graphs where each node u is connected to d nodes chosen independently at random. (If we ignore the directions of the edges, then each node has average degree $2d$ and neighborhood sets are only bigger.) We prove that the probability we do not get such an expander graph is strictly less than one. Hence, one must exist.

Event $E_{S,T}$: The graph G will not be a $\langle\alpha, \beta\rangle$-expander if there is some set S for which $|S| \leq \alpha n$ and $N(S) < \beta|S|$. Hence, for

each pair of sets S and T, with $|S| \leq \alpha n$ and $|T| < \beta|S|$, let $E_{S,T}$ denote the bad event that $N(S) \subseteq T$. Let us bound the probability of $E_{S,T}$ when we choose G randomly. Each node in S needs d neighbors for a total of $d|S|$ randomly chosen neighbors. The probability of a particular one of these landing in T is $\frac{|T|}{n}$. Because these edges are chosen independently, the probability of them all landing in T is $\left(\frac{|T|}{n}\right)^{d|S|}$.

Probability of Some Bad Event: The probability that G is not an expander is the probability that at least one of these bad events $E_{S,T}$ happens, which is at most the sum of the probabilities of these individual events.

$$\begin{aligned}\Pr[G \text{ not an expander}] &= \Pr[\text{At least one of the events } E_{S,T} \text{ occurs}] \leq \sum_{S,T} \Pr[E_{S,T}] \\ &= \sum_{(s \leq \alpha n)} \sum_{(S \mid |S|=s)} \sum_{(T \mid |T|=\beta s)} \Pr[E_{S,T}] = \sum_{s \leq \alpha n} \binom{n}{s}\binom{n}{\beta s}\left(\frac{|T|}{n}\right)^{d|S|}\end{aligned}$$

We now use the result that $\binom{n}{a} \leq \left(\frac{en}{a}\right)^a$.

$$\begin{aligned}\Pr[G \text{ not an expander}] &\leq \sum_{s \leq \alpha n} \left(\frac{en}{s}\right)^s \left(\frac{en}{\beta s}\right)^{\beta s} \left(\frac{\beta s}{n}\right)^{ds} = \sum_{s \leq \alpha n} \left[\left(\frac{en}{s}\right)\left(\frac{en}{\beta s}\right)^{\beta}\left(\frac{\beta s}{n}\right)^{d}\right]^s \\ &\leq \sum_{s \leq \alpha n} \left[\left(\frac{en}{\alpha n}\right)\left(\frac{en}{\beta \alpha n}\right)^{\beta}\left(\frac{\beta \alpha n}{n}\right)^{d}\right]^s = \sum_{s \leq \alpha n} \left[\frac{e^{\beta+1}}{\alpha} \cdot (\alpha\beta)^{d-\beta}\right]^s\end{aligned}$$

The requirement is that $\alpha\beta < 1$. Hence, if d is sufficiently big, $\left(d \geq \log\left(\frac{2e^{\beta+1}}{\alpha}\right) / \log\left(\frac{1}{\alpha\beta}\right) + \beta\right)$, then the bracketed amount is at most $\frac{1}{2}$.

$$\Pr[G \text{ not an expander}] \leq \sum_{s \leq \alpha n} \left[\frac{1}{2}\right]^s < 1$$

It follows that $\Pr[G \text{ is an expander}] > 0$, meaning that there exists at least one such G which is an expander.

Chapter 24

Entropy

Entropy is a hugely useful concept. We discuss it here in terms of the thermo dynamics, the expected number of bits needed to generate/specify a random object, and in compressing text.

Thermo Dynamics: The second law of thermo dynamics says that the Entropy of a closed system always increases. In physics, Entropy is a measure of how much usable energy there is. This amounts to how much disorder there is in a system. It is measured as some constant times the log of the number of *micro* states the system might be in given one knows the *macro* state. 100 years later, Shannon related Entropy to information theory. Because it takes $\log_2 N$ bits to specify one of N states, the Entropy of a system can be viewed as (a constant times) the number of bits of information needed to reveal the micro state. For example, if the macro state consist of a nicely ordered crystal, then there are few possible positions that the atoms may be in and it would take very few bits to reveal where they all are. On the other hand, if the macro state consist of a hot gas, each atom has some unknown location and velocity. It would then take a lot of bits to reveal all of this information.

Compressing Text: Suppose you had text consisting of a sequence of objects each from the set $\{Obj_1, \ldots, Obj_N\}$. Your task is to compress this text by allocating to each object Obj_i a code consisting of a short

bit string. If all n of the objects O_i are equally likely to appear in the text, then it makes sense to allocated each of them a code of length $\log n$. However, if some objects appear much more frequently then they should be allocated much shorter codes. One challenge with stringing together codes of different lengths is being able to uniquely decode what the original sequence of objects was. For example, you can not allocate Obj_1 the string 10, Obj_2 the string 11, and Obj_3 the string 1011, because then we would not know whether to decode 1011 as Obj_1Obj_2 or as Obj_3. It is sufficient to require that no code is the prefix of another code. This is best viewed a putting the objects Obj_i on the leaves of a binary tree. Label each left edge zero and each right edge one. The code for Obj_i will be the string of labels in the path from the root to the leaf it is on. One decodes 1011... by starting at the root and heading left or right down the tree as indicated by the bits. When one reaches a leaf, the object O_i at this leaf is outputted and one starts back at the root in order to decode the next object.

Code Length: The next task is to decide the optimal length I_i for each code. Focus for a moment only on the i^{th} object. We will argue that if it appears with probability p_i then optimally it should be allocated a code of length $I(p_i) = \log_1(\frac{1}{p_p})$. (Of course if this number is not an integer, then we might have to round it up a bit.) Here are two arguments for this. We have not considered all the other objects, but suppose they all had this same probability p_i of occurring. Then there would be $\frac{1}{p_i}$ objects and it would require $I(p_i) = \log_2(\frac{1}{p_i})$ bits to specify this object. The second argument is that if we allocate I_i bits to object Obj_i, then this object will be placed on a leaf of the binary tree at level I_i. In a full binary tree, there are 2^{I_i} nodes at this level. Hence, it is reasonable to say that that the object Obj_i has

"used up" $\frac{1}{2^{I_i}}$ of the tree. Given that it appears with probability p_i it should only "use up" a p_i fraction of the tree. This motivates setting I_i so that $\frac{1}{2^{I_i}} = p_i$. Solving gives that $I = \log_2(\frac{1}{p_i})$.

Building the Tree: Having decided to allocated code of length $I(p_i) = \log_2(\frac{1}{p_i})$ to object Obj_i, the next task is to allocated the codes themselves. It turns out, that there is always a way to build a binary tree with the objects Obj_i on its leaves so that objects Obj_i is at depth $\lceil \log_2(\frac{1}{p_i}) \rceil$.

Expected Code Length: Given that we allocated a code of length $I(p_i) = \log_2(\frac{1}{p_i})$ to object Obj_i, we can now compute the expected length of the code. Recall how expectation is computed.

$$H(\{p_i\}) = \text{Exp}_{i \in \{p_i\}} I(p_i) = \sum_i p_i I(p_i) = \sum_i p_i \log_2(\frac{1}{p_i})$$

It turns out that, given the probabilities p_i, this is the optimal expected length of the code. This then becomes a lower bound on how much the text can be compressed or a measure of the "information content" of the text. It is referred to as the *entropy* H of the probability distribution.

Entropy of a Probability Distribution: A probability distribution $\mathcal{D}$ on a set of objects/states $\{Obj_1, \ldots, Obj_N\}$ is defined by specifying for each state/object obj_i, the probability p_i of being in that state or of choosing that object. Entropy $H(\mathcal{D})$ measures the amount of randomness in a probability distribution. As seen above, it is the expected number bits that need to be communicated in order to specify which object was chosen. Within one or two, it is also the expected number of fair coins that need to be flipped to generate an object according to this distribution (See homework question).

24.1 Entropy Over View

Lemma 1 *The relationships between the entropies, joint entropies, conditional entropies, and the mutual information between three random variables X, Y, Z is equivalent to the relationships between the areas of three over lapping circles X, Y, Z.*

1. **Entropy:** $H(X) = \Sigma_x\, p(x)L(x)$, where $p(x)$ is short for $Pr(X = x)$ and $L(x)$ is short for $\log(1/p(x))$. $H(X) = area(X)$.

   Intuitively, $H(X)$ can be thought of as the expected length of shortest message to tell someone in the optimum way the value X happens to take on, where $L(x)$ is the length measured in bits to say $X = x$.

2. $0 \leq H(X) \leq \log_2(\#$ of different values).

3. $H(F(X)) \leq H(X)$ with equality when 1-1

4. **Joint Entropy:** $H(XY) = \Sigma_x \Sigma_y\, p(xy)L(xy)$, where $p(xy)$ is short for $Pr(X = x\ \&\ Y = y)$. $H(XY) = area(X \cup Y)$.

   Intuitively, It is the expected length of message to say both X & Y.

5. $H(XY) \leq H(X) + H(Y)$ with equality iff independent.

6. **Conditional Entropy:** $H(X|Y) = H(XY) - H(Y) = area(X \cap \overline{Y})$.

   Intuitively, $H(X|Y)$ is the expected length of message to tell you X after I have already told you Y.

7. $H(X|Y) \geq 0$ with equality iff Y determines X.

8. $H(X|Y) \leq H(X)$ with equality when independent.

9. $H(XY|Z) \leq H(X|Z) + H(Y|Z)$ with equality iff independent conditional on Z.

10. $H(F(X,Y)) \neq \Sigma_y H(F(X,y))$.

11. **Mutual Information:** $I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(XY) = area(X \cap Y)$.

    Intuitively, $I(X;Y)$ is the information that is common to both X and Y. It is the amount about X you learn from me telling you Y. Perhaps surprisingly, this is equal to the amount about Y you learn from me telling you X.

12. $I(X;Y) \geq 0$ with equality iff independent.

13. $I(X;Y) \leq H(X)$ with equality iff $X = Y$.

14. **Joint Mutual Information:** $I(XY;Z) = area((X \cup Y) \cap Z)$.

15. $I(XY;Z) \not\approx I(X;Z) + I(Y;Z)$ with $\geq$ if X and Y are independent.

16. $I(XY;Z) \leq I(X;Z) + H(Y)$.

17. $I(U;AR) \leq I(UR;A)$ when U and R are independent.

18. **Conditional Mutual Information:** $I((X;Y)|Z) = H(X|Z) - H(X|YZ) = area(X \cap Y \cap \overline{Z})$.

    Intuitively, it is the information common to X and Y after you have already told me Z. Or after you have already told me Z, it is the amount of additional information I learn about X from you telling me Y.

19. $0 \leq I((X;Y)|Z) \leq H(Y|Z) \leq H(Y)$.

20. $I((F(X,Y);Z)|Y) \leq I(X;Z)$.

21. **Strange Area:** $I(X;Y;Z) = Area(X \cap Y \cap Z) = I(X;Y) - I((X;Y)|Z) \not\geq 0$.

22. $I((X;Y;Z)|W) \geq 0$ when X and Z are conditionally independent given Y and W.

23. **Group Learning:** $\Sigma_j I((X_j;Z)|Y_j) \leq I((\langle X_1, X_2, \ldots, X_n \rangle ; Z)|\langle$ $H(Z)$ assuming for all disjoint subsets J and $J' \subseteq [n]$, X_J and $X_{J'}$ are independent conditional on $\langle Y_J, Y_{J'} \rangle$, and X_J is independent of $Y_{J'}$ conditional on Y_J.

    Intuitively, the sum of the amounts people in your class who came in with the knowledge X_i learned individually about their question Y_i from your lecture Z is at most that learned collectively.

24. **Possible Random Variables and Areas of Primitives:** Areas of circles X, Y, and Z correspond to random variables X, Y, and Z if and only if the area of each *dual primitive* must be positive, with the exception of $X \cap Y \cap Z$, which could be negative and for each pair of random variables the mutual information $I(X;Y) = X \cap Y$ must also be positive.

25. **Communication Complexity:** Let Π denote the transcript of a communication between two players with inputs X and Y and with private random bits. Consider a third player, Alice. Alice knowing both X and Y sends a message A consisting of m bits to the Y player (or to both of them.) Then the X and Y players have the conversation with transcript Π.

26. **Negative Area:**

    (a) $I(X;Y;\Pi) \geq 0$

    (b) $I(X;Y;A\Pi) \geq -m$.

    (c) $\Sigma_{k\in[K],j\in[n]} I(X_{\langle k,j\rangle}; Y_j; A\Pi) \geq -m$.
    This requires that conditioned on Y, the rows of X are independent.

It also requires that conditioned on Y, the bits $X_{\langle k,j\rangle}$ and $X_{\langle k,j'\rangle}$ are independent.
It also requires that conditioned on Y_{-j}, the bits $X_{\langle k,j\rangle}$ and Y_j are independent.
This is confusing. Just make all the bits independent.

(d) $\Sigma_{k\in[K],j\in[n]}\, I((X_{\langle k,j\rangle}; Y_j; A\Pi)|\left\langle\, X_{\langle *,-j\rangle}, Y_{-j}\,\right\rangle) \geq -nm$.
This requires that conditioned on Y, the rows of X are independent.

Where X is a matrix with $k \in [K], j \in [n], X_{\langle k,j\rangle}$ and Y_j a row.

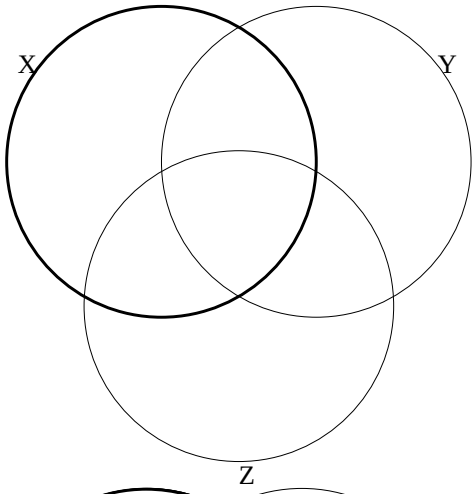
27. **What People Learn:** Suppose there is an eves dropper, Eve, who learns the conversation Π but knows neither X nor Y. By definition $I(XY;\Pi)$ is what she learns about the inputs $\left\langle\, X, Y\,\right\rangle$ from their conversation Π, $I(X;\Pi)$ is what she learns about the X-player's inputs X, and $I(Y;\Pi)$ about the Y-player's inputs Y. The X-player already knows X, and hence from Π the amount about Y that he learns is denoted $I((Y;\Pi)|X)$. Similarly, the Y-player learns $I((X;\Pi)|Y)$ about X.

- $I((Y;Z)|X) + I((X;Z)|Y) \geq I(XY;Z) \geq I(X;Z) + I(Y;Z)$ when X and Y are independent.
- $I((Y;\Pi)|X) + I((X;\Pi)|Y) \leq I(XY;\Pi) \leq I(X;\Pi) + I(Y;\Pi)$.
- Same with equality when X and Y are independent.
- $I((Y;A\Pi)|X) + I((X;A\Pi)|Y) \leq I(XY;A\Pi) + m$.

24.2 Entropy

Entropy

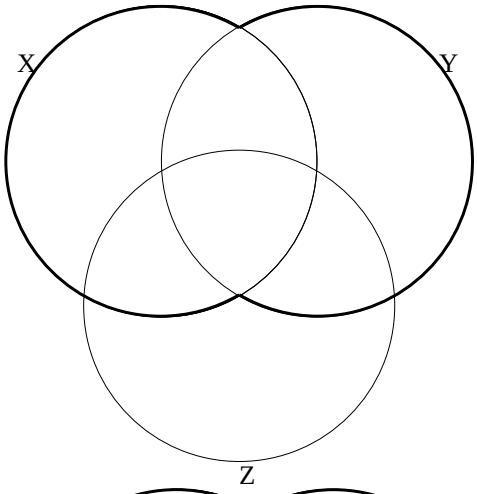
$H(X)$
$= \text{area}(X)$



Joint Entropy

$H(XY)$
$= \text{area}(X \cup Y)$

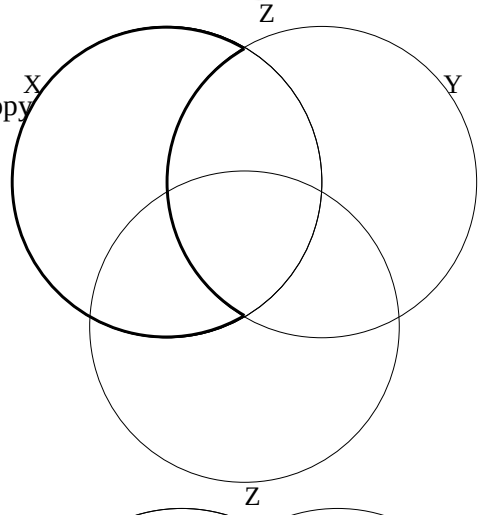
$\leq H(X)+H(Y)$
with equality when independent



Conditional Entropy

$H(X|Y)$
$= H(XY)-H(Y)$
$= \text{area}(X \cup Y) - \text{area}(Y)$

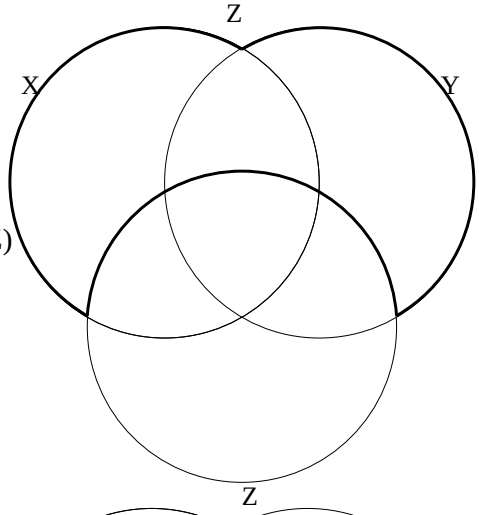
$\leq H(X)$



Joint Conditional Entropy

$H(XY|Z)$
$= H(XYZ)-H(Z)$

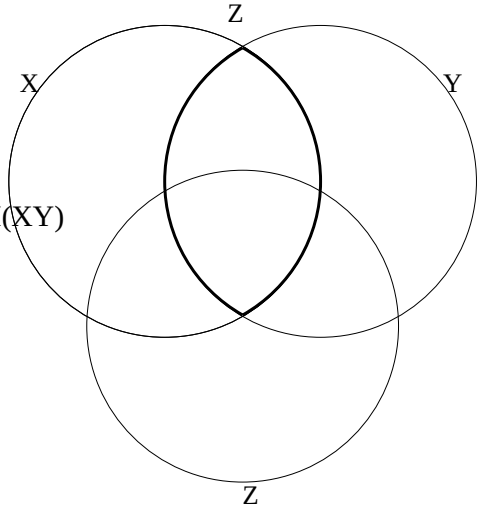
$\leq H(X|Z)+H(Y|Z)$
with equality when independent conditional on Z



Mutual Entropy

$I(X;Y)$
$= H(X)-H(X|Y)$
$= H(X)+H(Y)-H(XY)$
$= \text{area}(X \cap Y)$

≥ 0
with equality when independent

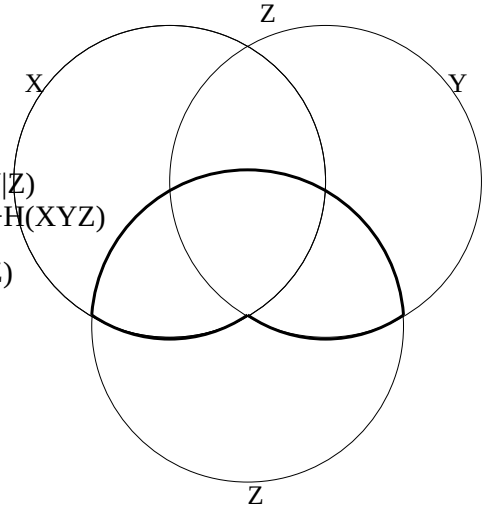


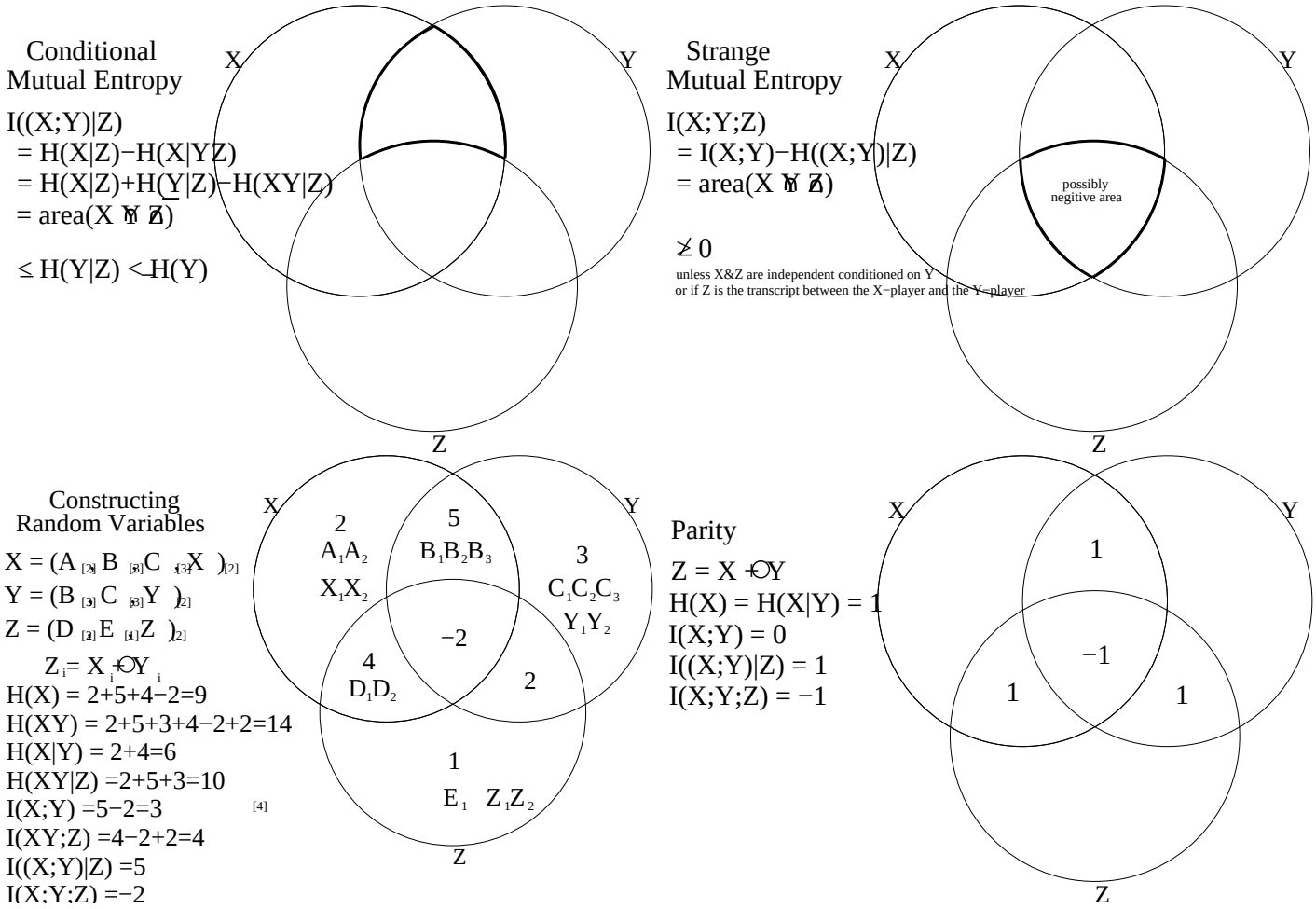
Joint Mutual Entropy

$I(XY;Z)$
$= H(XY)-H(XY|Z)$
$= H(XY)+H(Z)-H(XYZ)$

$\neq I(X;Z) + I(Y;Z)$
$\geq$ when x&y independent

$\leq I(X;Z) + H(Y)$





X, Y, and Z: X, Y, and Z can be thought of in three different ways. Being able to comfortably switch between these adds deeper understanding of the concepts and more tools for being able to prove results.

Random Variables: X, Y, and Z are defined to be three random variables. What objects/events/values these takes on does not matter, only how the probability is distributed between them. For example, X could be an apple with probability $\frac{1}{3}$ and an orange with probability $\frac{2}{3}$.

Information: A random variable adds a certain amount of uncertainty. Imagine that you are in the dark about which objects/events/values X happens to take on after all the deciding coins have been flipped, however, your friend does. Your friend has more *information* than you do. An alternative way of view-

ing X is not as a random variable but as the information that your friend has that you do not have.

Circles: Draw on the paper three intersecting circles and label them X, Y, and Z. Think of the circle X containing the information corresponding to knowing what value X takes on. The area outside of the circle X correspond to information that has nothing to do with the information X. We say that this information is *independent* of X.

Primitives and their Duals: Lets say that the *union* of a subset of the random variables is a *primitive*. For example, $(X \cup Y)$, often denoted $\langle X, Y \rangle$ and here simply XY. It will represent the information consisting of knowing both the value of X and of Y. Note if there are n random variables, there are 2^n of these. Now note that the three intersecting circles that you drew partition the page into $2^3 = 8$ separate areas. Lets call each of these the *dual primitives*. We specify a dual primitive by specifying for each random variable whether we are considering the area inside or outside of its corresponding circle. For example $X \cap \overline{Y} \cap Z$, denotes the area that in X, in Z, and outside of Y. There are 2^n of these. Given the area, measured in say cm^2, of each of these dual primitives, one can compute the area of each primitive. For example, when there are two circles, the area of $X \cup Y$ = the area of $X \cap \overline{Y}$ + the area of $X \cap Y$ + the area of $\overline{X} \cap Y$. Similarly, if you know the area of each primitive, one can compute the area of each dual primitive, because there are 2^n linear equations and 2^n unknowns.

NOT Venn Diagrams: Be very careful not to confuse these circles with Venn diagrams. In a Venn diagram, X, Y, and Z must be random variables with taking on either *true* or *false*.

Inside the circle X represent all results of the coin flips that leads to X being true and outside represent all those for which X is false. The area of the circle is the probability that X is true.

Entropy $H(X)$ (Proof of Lm1.1): Just as there are three ways of understanding the random variable X, there are corresponding ways of understanding the *entropy* $H(X)$ of X.

Random Variables: $H(X)$ is said to be the *entropy of* X.

Computationally: It is formally defined as $H(X) = \Sigma_x\, p(x)L(x)$, where $p(x)$ is short for $Pr(X = x)$ and $L(x)$ is short for $\log(1/p(x))$.

Information: $H(X)$ can be thought of as the expected length of shortest message to tell someone in the optimum way the value X happens to take on, where $L(x)$ is the length measured in bits to say $X = x$.

Generating X: If you were to write a program that flips as few coins as possible in order to determine which value X should take, then $H(X)$ would be the expected number of coins that would need be flipped, where $L(x)$ is the number flipped when the program decides that $X = x$.

Circles: $H(X)$ can also be viewed as the area of the circle representing X.

$0 \leq H(X) \leq \log_2$(# of different values) (Proof of Lm1.2 One can identify each n different objects, using a binary label containing $\log_2(n)$ bits. Given this is one way to communicate X, the optimal way takes at most this many bits.

$H(F(X)) \leq H(X)$ with Equality when 1-1 (Proof of Lm

Here F is a function that maps each objects/events/values that X takes to some other object/event/value. For each of these new objects f, there is a probability that $F(X,Y) = f$. Hence, F becomes a random variable in its own right. Remember that the which objects X takes on does effect the entropy $H(X)$ of X, only how the probability is distributed between them. Hence, if X takes on the objects apple, orange, and pair, and f maps these to 1, 2, and 3, then the entropy does not change. However, if f collapse apple and orange to the same value 1, then the entropy goes down.

Proof: The key observation is that for each x, $\Pr(F(X) = F(x)) \geq \Pr(X = x)$, because when ever $X = x$, we have that $F(X) = F(x)$, but it may be that $F(X) = F(x') = F(x)$ when $X = x' \neq x$. From this we bound the entropy. $H(F(X)) = \Sigma_f \Pr(F(X,Y) = f) \log(1/\Pr(F(X) = f) = \Sigma_f [\Sigma_x p_x(\text{whether } F(x)=f)] \log(1/\Pr(F(X) = f) = \Sigma_x p_x \log(1/\Pr(F(X)=F(x)) \leq \Sigma_x p_x \log(1/\Pr(X=x)$.

Joint Entropy $H(XY)$ (Proof of Lm1.4): (Often written $H(X,Y)$ or $H(\langle X,Y\rangle)$.

Random Variables: If X and Y are random variables then $\langle X,Y\rangle$ is the *joint* random variable telling you both the value of X and of Y. The joint entropy, $H(XY)$, is to defined to be the entropy of $\langle X,Y\rangle$.

Computationally: It is formally defined as $H(XY) = \Sigma_x \Sigma_y p(xy)L(xy)$, where $p(xy)$ is short for $Pr(X = x \;\&\; Y = y)$.

Information: It follows that $H(XY)$ is the expected length of message to say both X & Y.

Circles: $H(XY)$ is the area of $X \cup Y$.

$H(XY) \leq H(X) + H(Y)$ with Equality iff Independent

Intuition: Intuitively, this is true because one could say both X and Y by saying X and then saying Y, but perhaps one could save time by saying them together if some of the information overlaps. But if X & Y are independent, then no information overlaps and $H(XY) = H(X) + H(Y)$.

Proof: With independence the proof is as follows.

$L(xy) = \log(1/p(xy))$, which by independence is $\log(1/(p(x)p(y))) = L(x) + L(y)$. This gives $H(XY) = \Sigma_x \Sigma_y p(xy)L(xy) = \Sigma_x \Sigma_y p(xy)L(x) + \Sigma_x \Sigma_y p(xy)L(y) = \Sigma_x [\Sigma_y p(xy)] L(x) + \Sigma_y [\Sigma_x p(xy)] L(y) = \Sigma_x p(x)L(x) + \Sigma_y p(y)L(y) = H(x) + H(y)$.

The proof that without independence $H(XY) \leq H(X)+H(Y)$ is harder. Note that the terms for which $L(xy) > L(x) + L(y)$ have their weight $p(xy)$ larger because $p(xy) > p(x)p(y)$. Similarly the terms for which $L(xy) < L(x) + L(y)$ have their weight $p(xy)$ smaller because $p(xy) < p(x)p(y)$.

Conditional Entropy $H(X|Y)$ (Proof of Lm1.6):

Random Variables: $H(X|Y)$ is defined to be the entropy of X *conditional on* Y.

Computationally: It is formally defined as $H(X|Y) = \Sigma_y p(y)H(X|y)$, but this is not so intuitive.

Information: Intuitively, $H(X|Y)$ is the expected length of message to tell you X after I have already told you Y.

Circles: Pictorially, $H(X|Y)$ is the area of $X - Y = X \cap \overline{Y}$. This is the area of $X \cup Y$ minus the area of Y.

$H(X|Y) = H(XY) - H(Y)$: This seems to be a more intuitive and a more useful definition of mutual information.

Intuition: Thinking of mutual information as areas of area of $X \cup Y$ minus the area of Y leads us to understand that $H(X|Y) = H(XY) - H(Y)$ is true.

Primitives: I like this definition because it is wrt the "primitives" $H(X)$, $H(Y)$ and $H(XY)$.

Conditional Probabilities: I also like it because it looks like $p(x|y) = \frac{p(xy)}{p(y)}$ when you take the log of both sides.

Proof: The proof of $H(X|Y) = H(XY) - H(Y)$ is as follows.
$H(X|Y) = \Sigma_y\, p(y) H(X|y) = \Sigma_y\, p(y) \left[\Sigma_x\, p(x|y) L(x|y)\right]$
$= \Sigma_y\, p(y) \left[\Sigma_x\, \frac{p(xy)}{p(y)} \log(\frac{p(y)}{p(xy)})\right] = \Sigma_x\, \Sigma_y\, p(xy) \left[L(xy) - L(y)\right]$
$= \left[\Sigma_x\, \Sigma_y\, p(xy) L(xy)\right] - \left[\Sigma_x\, \Sigma_y\, p(xy) L(y)\right] = H(XY) - \Sigma_y \left[\Sigma_x\, p(xy)\right] L(y)$
$= H(XY) - \Sigma_y\, p(y) L(y) = H(XY) - \Sigma_y\, p(y) L(y) = H(XY) - H(Y)$.

$H(X|Y) \geq 0$ with Equality iff Y Determines X (Proof

Intuition: After telling you Y, either you know X in which case $H(X|Y) = 0$ or I need to tell you more.

Proof: The proof is easy using our new definition, $H(X|Y) = H(XY) - H(Y)$ which is positive because surely it take more to tell you X and Y than to simply tell you Y.

$H(X|Y) \leq H(X)$ with Equality when Independent (Pr

Intuition: The intuition is that it can only be easier to tell you X after I have only told you Y.

Proof: The proof is easy using our new definition, $H(X|Y) = H(XY) - H(Y) \leq [H(X) + H(Y)] - H(Y) = H(X)$.

$H(XY|Z) \leq H(X|Z) + H(Y|Z)$ with Equality iff Indep

Intuition: After I have told you Z, it is no harder to tell you X and Y together than to tell you each separately.

Proof: The proof of this is hard like that for $H(XY) \leq H(X) + H(Y)$.

$H(F(X,Y)) \neq \Sigma_y H(F(X,y))$ (Proof of Lm1.10): It is natural to assume that because entropy is a weighted sum that one can decompose it into a weighted sum. However, it is not this easy. $\Sigma_y H(F(X,y)) = \Sigma_y H(F(X,Y)|Y = y) = H(F(X,Y)|Y) \neq H(F(X,Y))$.

Mutual Information $I(X;Y)$ (Proof of Lm1.11):

Random Variables: $I(X;Y)$ is said to be the *mutual information* between X and Y.

Information: $I(X;Y)$ is the information that is common to both X and Y.
It is the amount about X you learn from me telling you Y.
Perhaps surprisingly, this is equal to the amount about Y you learn from me telling you X.

Computationally: It is formally defined as $I(X;Y) = H(X) - H(X|Y)$,
namely how much less do I have to tell you to tell you X after I have told you Y.

Primitives: Decomposed it into primitives gives $I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(XY)$.

Circles: $I(X;Y)$ is the area of $X \cap Y$. This is the area of X plus that of Y minus that of $X \cup Y$.

$\langle X;Y \rangle$ is Not a Random Variable: When seeing $H(X;Y)$, it is tempting to think of $\langle X;Y \rangle$ as a random variable and/or information to be be communicated. This is not the case and can lead to faulty intuition at times. The information $\langle X;Y \rangle$ common between X and Y is something that may be hard to communicate separate from communicating X and Y. It is not a random variable to communication in its own right. Maybe this is why they use the letter I in $I(X;Y)$ instead of an H as in $H(X;Y)$.

$I(X;Y) \geq 0$ with Equality iff Independent (Proof of Lm

Intuition: $I(X;Y) \geq 0$ is true because the amount of common information cant be negative. Clearly if X and Y are independent then $I(X;Y) = 0$, because they contain no information about each other.

Proof: $I(X;Y) \geq 0$ follows from our new definition $I(X;Y) =$ $H(X) + H(Y) - H(XY)$ and that $H(XY) \leq H(X) + H(Y)$ with equality iff independent.

$I(X;Y) \leq H(X)$ with Equality iff $X = Y$ (Proof of Lm

Intuition: $I(X;Y) \geq 0$ is true because the amount of common information cant be negative. Clearly if X and Y are independent then $I(X;Y) = 0$, because they contain no information about each other.

Proof: $H(X) - I(X;Y) = H(X) - [H(X) - H(X|Y)] =$ $H(X|Y) \geq 0$.

Joint Mutual Information $I(XY;Z)$ (Proof of Lm1.14):

When I tell you Z, what does this tell you about the joint entropy of X and Y?

$I(XY;Z) \not\approx I(X;Z) + I(Y;Z)$ (Proof of Lm1.15): When understanding a subject, it is important to understand not only what is true but also what intuitively one might think is true but is not. Despite ones intuition, no direct comparison can be made between $I(XY;Z)$ and $I(X;Z) + I(Y;Z)$.

$I(XY;Z) \not\leq I(X;Z) + I(Y;Z)$:

False Intuition: One might first guess that this would be similar to $H(XY) \leq H(X) + H(Y)$, namely that the amount you learn about XY from Z is at most the you learn about X plus the amount you learn about Y.

Example: Consider the example in which $X = Y$ and $I(X;Z) > 0$. Then $I(XY;Z) = I(X;Z) = I(Y;Z)$. It follows that for this example $I(XY;Z) < 2I(XY;Z) = I(X;Z) + I(Y;Z)$.

$I(XY;Z) \not\geq I(X;Z) + I(Y;Z)$:

False Intuition: On the other hand, one could argue the opposite. You learn at least as much about Z from X and Y than you learn about Z from each of them separately.

Example: Consider the example $Z = X \oplus Y$, where X and Y are independent boolean variables. Here knowing Z tells you nothing about X and similarly nothing about Y. However, knowing Z tells you a full bit of information about $\langle X, Y \rangle$, namely their parity. It follows that for this example $1 = I(XY;Z) > I(X;Z) + I(Y;Z) = 0$.

$I(XY;Z) \geq I(X;Z) + I(Y;Z)$ if X and Y are indep

See the intuition and example above.

Proof: The proof is as follows. $I(XY;Z) = H(XY) - H(XY|Z)$. Because of independence, we have that $H(XY) = H(X) + H(Y)$. However, we can only be sure that $H(XY|Z) \leq H(X|Z) + H(Y|Z)$ unless, X and Y are independent conditional on Z. It follows that $I(XY;Z) \geq [H(X) - H(X|Z)] + [H(Y) - H(Y|Z)] = I(X;Z) + I(Y;Z)$.

Key: This is key in proving lower bound in communication complexity.

$I(XY;Z) \leq I(X;Z) + H(Y)$ (Proof of Lm1.16): This is similar in nature to the previous comparison, but it is true.

Intuition: The intuition is as follows. Suppose someone knows X and from this can deduce $I(X;Z)$ about Z. Then suppose someone tells him Y. Now he knows X and Z. The amount that he can now deduce about Z is denoted $I(XY;Z)$. Surely this is at most the about $I(X;Z)$ that he could deduce before plus the number of bits $H(Y)$ needed to communicate Y.

Proof: This can be proved as follows. Look at the 2^3 dual primitive formed from the intersections of the three circles for X, Y, and Z. In each separate dual primitive area put a +1 when within the area $X \cap Z$ for $I(X;Z)$ and another +1 when within the area Z. Similarly, put a -1 when within each dual primitive area within $XY \cap Z$ for $I(XY;Z)$. Summing these gives a zero in every dual primitive area except for three, which are the union of $X \cap Y$ and $Y - (X \cup Z)$. These two areas are equal to $I(X;Y)$ and $H(Y|XZ)$. All this proves that $RHS - LHS = I(X;Z) + H(Y) - I(XY;Z) = I(X;Y) + H(Y|XZ)$. One could prove this more formally by expanding each into the prim-

itives and making sure they all cancel. We also know that both $I(X;Y)$ and $H(Y|XZ)$ are positive. The statement follows.

$I(U;AR) \leq I(UR;A)$ when U and R are independent (

$I(U;AR) = I(U;A|R) + I(U;R) = I(U;A|R) + 0 \leq$ $I(U;A|R) + I(A;R) = I(UR;A)$. ■

Conditional Mutual Information $I((X;Y)|Z)$ (Proof of Lm

It is often written as $I(X;Y|Z)$ but its is parsed as $I((X;Y)|Z)$.

Information: $I((X;Y)|Z)$ is the information common to X and Y after you have already told me Z. Or after you have already told me Z, $I((X;Y)|Z)$ is the amount of additional information I learn about X from you telling me Y.

Computationally: $I((X;Y)|Z) = H(X|Z) - H(X|YZ)$
$= H(X|Z) + H(Y|Z) + H(XY|Z)$
$= H(XZ) + H(YZ) - H(Z) - H(XYZ)$.

Circles: $I((X;Y)|Z)$ is the area of $(X \cap Y) - Z$ which is the single dual primitive $X \cap Y \cap \overline{Z}$.

$0 \leq I((X;Y)|Z) \leq H(Y|Z) \leq H(Y)$ (Proof of Lm1.19):

All of these are reasonable and useful things that are true but need to be proved.

Proof of $I((X;Y)|Z) \geq 0$: By definition $I((X;Y)|Z) =$ $H(X|Z) + H(Y|Z) + H(XY|Z)$. But we have already seen that $H(XY|Z) \leq H(X|Z) + H(Y|Z)$.

Proof of $I((X;Y)|Z) \leq H(Y|Z)$: $H(Y|Z) -$ $I((X;Y)|Z) = H(Y|Z) - [H(Y|Z) - H(Y|XZ)] =$ $H(Y|XZ) \geq 0$.

Proof of $H(Y|Z) \leq H(Y)$: $H(Y) - H(Y|Z) = I(Y;Z) \geq$ 0.

$I((F(X,Y);Z)|Y) \leq I(X;Z)$ (Proof of Lm1.20): The intuition is that if I tell you the value of Y, then $F(X,Y)$ simply becomes a function $F_y(X) = F(X,y)$ dependent only on X. As we have seen above $H(F_y(X)) \leq H(X)$. Similarly, $I((F_y(X);Z) \leq I(X;Z)$.

Proof: Could write one ????

Strange Area $I(X;Y;Z) = Area(X \cap Y \cap Z) \not\geq 0$ (Proof o

The dual primitive $X \cap Y \cap Z$ does not really have a defined meaning using entropy, conditional entropy, and mutual entropy. We have seen that $I(X;Y) = Area(X \cap Y)$ is the information that is mutual between X and Y, hence for ease of notation, let us denote the area of $X \cap Y \cap Z$ by $I(X;Y;Z)$. For what ever it means this would be the information that is "mutual between all three." The problem with this area is that for some random variables it is positive and for some it is negative. This should be understood because it is the source why our intuition about entropy sometimes goes wrong.

$I(X;Y;Z) = I(X;Y) - I((X;Y)|Z) \not\geq 0$: The question $I(X;Y)$ vs $I((X;Y)|Z)$ is whether or not there can be more common information between X and Y after I tell you Z.

False Intuition: We already have seen $I(W;Z) = H(W) - H(W|Z) \geq 0$. Hence, it would be natural to generalize to $I(X;Y;Z) = I((X;Y);Z) = H((X;Y)) - H((X;Y)|Z) \geq 0$. However, we will see that this is not true. It is not directly true because $\langle X;Y \rangle$ is not a random variable in its own right.

Circles: Lets try to expand $I(X;Y) - I((X;Y)|Z) =$ $[H(X) + H(Y) - H(XY)] - [H(XZ) + H(YZ) - H(Z) - H(X$ Look at the three intersecting circles. In each dual primitive area put a +1 when the area is added and a -1 when it is

subtracted. Summing these gives a zero every where except for a one in the intersection dual primitive area $X \cap Y \cap Z$. This concurs with the intuition that this is the information that is common between all three of X, Y and Z, namely $I(X;Y;Z) = I(X;Y) - I((X;Y)|Z)$.

Counter Example: This same example often comes up. Let Y and Z be independent boolean random variables. Let $X = Y \oplus Z$.
Note $I(X;Y) = 0$ because X and Y are independent.
However, if I tell you Z, then you know X from Y and hence $I((X;Y)|Z) = H(Y) = 1$.
Combining all the entropies gives $I((X;Y)|Z) = H(XZ) + H(YZ) - H(Z) - H(XYZ) = 2 + 2 - 1 - 2 = 1$.
In this case, $1 = I((X;Y)|Z) > I(X;Y) = 0$.

Negative Area: The interesting thing is that $X = Y \oplus Z$ gives an example where $RHS - LHS = I(X;Y) - I((X;Y)|Z)$ is negative giving that the "area" of $X \cap Y \cap Z$ is negative. In contrast, we can prove that every other dual primitive area must be positive. I don't know if this is a coincidence or not, but pictorially this is reasonable. If two circles over lap, then their intersection should be positive. If they don't over lap, then their intersection should be zero. If three circles over lap, then their intersection should be positive. If they don't over lap but lie in say a line then their intersection should be zero. But what happens if they form a circle, with X overlapping with Y, Y with Z and Z with X, and with space in the middle of the circle. This is like having $X \cap Y \cap Z$ have negative area.

$I((X;Y;Z)|W) \geq 0$ when X and Z are conditionally in

Positive Area: $I(X;Y;Z)$ is the area of the dual primitive $X \cap Y \cap Z$. It is counter intuitive for this area to be negative. We claim, in fact, that most counter intuitive things about mutual entropy arise from the fact that this area can be negative. Hence, it is interesting to look at conditions in which it is not. Conditioning it on another variable W, just makes the result more general. Assuming W is a constant removes mention of it.

$I((X;Y)|Z) \leq I(X;Y)$ when X and Z are condition This is a restatement of the above because by definition $I(X;Y;Z) = I(X;Y) - I((X;Y)|Z)$. This was listed in the [YJLS]. I am not sure how they used, it but we will find it very useful.

Proof: By assumption, X and Z are conditionally independent given Y and W. Formally this means $I((X;Z)|YW) = 0$. We have not considered the intersections of four random variables. But we can think of $U = YW$ as corresponding to $U = Y \cup W$. Above we defined $I((X;Z)|U) = area(X \cap Z \cap \overline{U}) = area(X \cap Z \cap \overline{Y \cup W}) + area(X \cap Z \cap \overline{Y} \cap \overline{W}) = 0$.
What is always true is that mutual information $I((X;Z)|W) = area(X \cap Z \cap \overline{W}) \geq 0$.
It follows that $I((X;Y;Z)|W) = area(X \cap Z \cap Y \cap \overline{W}) = area(X \cap Z \cap \overline{W}) - area(X \cap Z \cap Y \cap \overline{W}) = I((X;Z)|W) - 0 \geq 0$.

Group Learning (Proof of Lm1.23): $\Sigma_j I((X_j;Z)|Y_j) \leq I((\langle X_1, X_2, \ldots, X_n \rangle ; Z)| \langle Y_1, Y_2, \ldots, Y_n \rangle) \leq H(Z)$ assuming for all disjoint subsets J and $J' \subseteq [n]$, X_J and $X_{J'}$ are independent conditional on $\langle Y_J, Y_{J'} \rangle$, and X_J is independent of $Y_{J'}$ conditional on

Y_J.

Intuition: Suppose you are teaching a class to n people. For each $j \in [n]$, let Y_j denote the information that the j^{th} person knows before the lecture. Let X_j denote the information that he personally wants to learn, hopefully from the lecture. Let Z denote the information taught at the lecture. Then $I((X_j; Z)|Y_j)$ denotes how much the j^{th} person learns during the lecture about his personal question and $I((\langle X_1, X_2, \ldots, X_n\rangle ; Z)|\langle Y_1, Y_2, \ldots, Y_n\rangle)$ denotes how much the class collectively learns about their collective questions. Requiring that X_J and $X_{J'}$ are independent conditional on $\langle Y_J, Y_{J'}\rangle$ asserts that for any disjoint subsets of people $J \neq J' \in [n]$, given their combined knowledge, their questions are independent of each other. Requiring that X_J is independent of $Y_{J'}$ conditional on Y_J asserts that given what the J of people know, the other people's previous knowledge will not help him with his personal question. The conclusion is that the sum of the amounts learned individually is at most that learned collectively.

Proof: It is only necessary to prove it for two people, then the result for n people follows by induction.

(X_1 and X_2 are independent conditional on $\langle Y_1, Y_2\rangle$) thought of as areas of circles translates into $area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2) = 0$.

($I((X_1; X_2)|Y_1Y_2Z) \geq 0$) translates into $area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap \overline{Z}) \geq 0$.

Combining the last two statements gives that $area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) = area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2) - area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap \overline{Z}) \leq 0$. Call this statement (1).

(X_2 is independent of Y_1 given Y_2) translates into $area(X_2 \cap Y_1 \cap \overline{Y}_2) = 0$.

$(I((X_2;Y_1)|Y_2Z) \geq 0)$ translates into $area(X_2 \cap Y_1 \cap \overline{Y}_2 \cap \overline{Z}) \geq 0$.
Combining the last two statements gives that $area(X_2 \cap Y_1 \cap \overline{Y}_2 \cap Z) = area(X_2 \cap Y_1 \cap \overline{Y}_2) - area(X_2 \cap Y_1 \cap \overline{Y}_2 \cap \overline{Z}) \leq 0$.
We use this result to bound the following. $I((X_2;Z)|Y_2) = area(X_2 \cap \overline{Y}_2 \cap Z) = area(X_2 \cap Y_1 \cap \overline{Y}_2 \cap Z) + area(X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) \leq area(X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) = area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) + area(\overline{X}_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z)$. Call this statement (2).
By symmetry of statement (2), get the following statement (3) that $I((X_1;Z)|Y_1) \leq area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) + area(X_1 \cap \overline{X}_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z)$.
$I((\langle X_1, X_2 \rangle ; Z)|\langle Y_1, Y_2 \rangle) = area((X_1 \cup X_2) \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) = area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) + area(\overline{X}_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) + area(X_1 \cap \overline{X}_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z)$. Call this statement (4).
Combining statements (2,3,4) gives that $I((X_1;Z)|Y_1) + I((X_2;Z)|Y_2) - I((\langle X_1, X_2 \rangle ; Z)|\langle Y_1, Y_2 \rangle) \leq area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z)$. Then by statement (1), this less than or equal to zero.
■

Possible Random Variables and Areas of Primitives (Proof

In trying to prove lemmas about entropy, one needs to know the range of possible interactions between random variables. Recall that the n circles corresponding to n random variables partition the paper into 2^n dual primitives like $X \cap \overline{Y} \cap Z$. Suppose you are told the area of each of these dual primitives. One might ask, for which such values can one construct random variables such that the entropy, conditional entropy, and mutual entropy are all given by these areas. We have answered this question given only three random variables.

Requirements on Dual Primitive Areas: The first requirement is that we have (or could) prove that the area of each dual

primitive must be positive, with the exception of $X \cap Y \cap Z$, which we affectionately called $I(X;Y;Z)$. The second requirement is that for each pair of random variables the mutual information $I(X;Y) = X \cap Y$ must also be positive. These are the only requirement for constructing random variables X, Y, and Z with these values.

Proof: For convenience of the argument, we will assume that all of the require areas are integers. (It is trickier but still doable when they are arbitrary reals.) To begin, if $X \cap Y \cap Z$ needs to have negative area, then let $a = |area(X \cap Y \cap Z)|$. Let $U = \langle u_1, \ldots, u_a \rangle$ and $V = \langle v_1, \ldots, v_a \rangle$ be the results of $2a$ independent random coin flips. Let $W = \langle w_1, \ldots, w_a \rangle$ such at $w_i = u_i \oplus v_i$. We have already see that the area of the dual primary $I(u_i; v_i; z_i) = u_i \cap v_i \cap u_i$ is -1 and that the area of $u_i \cap v_i \cap \overline{u}_i$, $u_i \cap \overline{v}_i \cap u_i$, and $\overline{u}_i \cap v_i \cap u_i$ are each 1. Hence, these areas for U, V, and W will be $-a$ and a. Include U in the description of X, V in Y, W in Z. This accounts for $-a$ and a of the areas of these primary primitives for X, Y, and Z. Subtract off these amounts from the required areas of the primitives. All the new requirements on the primitives are now positive. Subtracting $-a = area(X \cap Y \cap Z)$ from the required area of $X \cap Y \cap Z$ makes its new required area zero. The new required area for $X \cap Y \cap \overline{Z}$ is equal to the original required area for it minus a, which is $area(X \cap Y \cap Z) + area(X \cap Y \cap \overline{Z}) = I(X;Y) \geq 0$.

We complete the construction as follows. For each dual primitive whose new required area is b define a new random variable consisting of the results of b new independent coin flips. Include this information in each random variable that the dual primitive is in. For example, let $b_{X \cap Y \cap \overline{Z}}$ be the new required area for the dual

primary $X \cap Y \cap \overline{Z}$. Let $U_{X \cap Y \cap \overline{Z}}$ consist of $b_{X \cap Y \cap \overline{Z}}$ independent coin flips. Add $U_{X \cap Y \cap \overline{Z}}$ to X and Y, but not to Z. It follows that X, Y, and Z have the entropy, conditional entropy, and mutual entropy as given by the required areas of the dual primitives.

24.3 Predictability

In this section we introduce Russell Impagliazzo's idea of *Predictability*.

- J. Edmonds, R. Impagliazzo, S. Rudich, and J. Sgall, "Communication Complexity Towards Lower Bounds on Circuit Depth," *Journal of Computational Complexity,* 10: pp 210-246, 2001. Previously in *FOCS, Symp. Foundations of Computer Science*, pp. 249-257, 1991.

For example, suppose over a sequence of communication bits, the A-Player tells the B-Player, "I am not telling you any of my values, but I will tell you that if my coin flip is heads, then all k bits in my vector are zero. On the other hand, if this flip is tails, then I reveal no information about this vector." The question now is whether the B-Player is considered to know k or zero bits about this vector. A useful information measure for many applications is Entropy. Because half the time k bits about the vector are revealed and half the time 0 bits are revealed, Entropy measures the number of bits revealed as the average $(k + 0)/2 = k/2$. Our adversary, however, wants to be more cautious by assuming that the B-Player knows more than this. We define a measure of "predictability" to be the probability of guessing the value. If the B-Player completely knows the vector then the probability of guessing it is 1 and if he knows nothing about it, then the probability is 2^{-k}. The measure is the average of these, $(1+2^{-k})/2 \approx 1/2$. A predictability of $1/2$ is then interpreted to

mean that the B-Player knows everything but "1 bit" about this vector. The next section formally defines this measure.

Def^n 24.3.1 *Let $v \in [k]^\pi$ a vector be a vector indexed by enteries in π and values in $[1..k]$ and let S be a set of such vectors. For every subset of indices $\rho \subseteq \pi$, let $Proj(v, \rho) \in [k]^\rho$ be the sub-vector of v indexed by the indices in ρ and let $Proj(S, \rho) = \{Proj(v, \rho) \mid v \in S\}$ be the* **projection** *of S onto ρ. A function R from $Proj(S, \rho)$ to S is called an* **extension function** *if for all $w \in Proj(S, \rho)$, $Proj(R(w), \rho) = w$.*

Suppose an input v specifies for each index $i \in \pi$, a value $v_i = Proj(v, i) \in [k]$. Then there are $k^{|\pi|}$ possible inputs v. Suppose someone restricts this set of possible inputs to a set of size $|S| \geq \left(\frac{r}{k}\right)^\ell \times k^{|\pi|}$. We will interpret this as them revealing some $t = \ell \log(k/r)$ bits about the input. An interesting question is, for $i \in \pi$, how many of these bits were communicated "about the i^{th} element" conditioned on knowing the other elements? Since the actual bits communicated could depend on all the elements, this is not a clear-cut issue. Our measure is computed as follows. Choose a random vector $w \in Proj(S, \pi - i)$ giving values of all the elements other than i. The set $\{v_i \in [k] \mid \langle w, v_i \rangle \in S\}$ $\equiv \{v \in s|\ Proj(v, \pi - i) = w\}$ is the set of values for the i^{th} element (or full vectors) consistent with our chosen values w for the other elements. We define the unpredictability of v_i to be the expected number of such choices.

Def^n 24.3.2 *The* **unpredictability** *of the i^{th} element in S is $UnPred_i(S) = \frac{|S|}{|Proj(S, \pi - i)|}$.* [1]

[1] This definition of *unpredictability* is one over the the definition of *predictability* defined in [EIRS]. They also give a second equivalent definition. Also our $Proj(v, \rho)$ is their $Proj(v, \pi - \rho)$.

If the i^{th} element v_i of $v \in S$ is fixed as a function of the other elements, then $UnPred_i(S) = 1$. If this element is completely undetermined, then $UnPred_i(S) = k$. If $UnPred_i(S) = r$, we can think of $t = \log(k/r)$ as the "number of bits known about element i", since if S is the set of inputs consistent with this number of independent bits communicated about v_i, then $UnPred_i(S) = \frac{k}{2^t} = r$.

Suppose $t = \ell \log(k/r)$ bits have been communicated about the vectors v in S, i.e., $|S| = 2^{-t} \cdot k^{|\pi|}$. A natural property to want is that at most ℓ elements of v can have more than $\frac{t}{\ell} = \log(k/r)$ bits "revealed about it". The exemplary counter example is the following. Suppose that the sum of the elements over the field $[k]$ was revealed. This requires only $\log(k)$ bits to be communicated. On one hand, it feel like nothing has been revealed about any one element because each still can uniformly take on any value. On the other hand, each element has been completely revealed, conditioned on knowing the other elements. This gives that for each $i \in \pi$, $UnPred_i(S) = 1$, implying that $\log(k)$ bits have been communicated "about each of the elements". We will get around this problem as follows. When an element $i \in \pi$ becomes highly predictable in S, we start ignoring it by considering only the possible settings of the other elements $Proj(S, \pi - i)$ in place of S. The value of element i is fixed as a function of the other elements using an extension function $R(w)$ as defined in Definition 24.3.1. In our previous example, if any one of element v_i is fixed to be $v_i = sum - \Sigma_{j \neq i} v_j$, then no information at all is known about the remaining elements. The following lemma then gives us what we wanted, that if at most $t = \ell \log(k/r)$ bits have been revealed about S, then there exists a set of at most ℓ elements such that, if we fixed them in this way, then no more than $\frac{t}{\ell} = \log(k/r)$ bits have been "revealed about" any of the other elements and hence there are still $\frac{k}{2^{t/\ell}} = r$ values left for each.

Lemma 2 *[Lemma 4.6 in [EIRS]] Let* $|S| \geq \left(\frac{r}{k}\right)^{\ell} \times k^{|\pi|}$ *and* $|\pi| \geq \ell >$ 0. *Then there exists a subset* $\sigma \subseteq \pi$ *of at most* ℓ *of elements such that if we* reveal *these, then each of the unrevealed elements is still highly unpredictable, namely* $\forall i \in \pi - \sigma$, $UnPred_i(Proj(S, \pi - \sigma)) > r$.

Proof:

Initially, let $\sigma = \emptyset$. We will keep adding indices to σ, maintaining the property that

$$\frac{|Proj(S, \pi - \sigma)|}{k^{|\pi - \sigma|}} \geq \left(\frac{r}{k}\right)^{\ell - |\sigma|}.$$

Clearly this is true for $\sigma = \emptyset$, because $Proj(S, \pi) = S$. Now assume for $\sigma \subseteq \pi$ the property holds and that there is an index $i \in \pi - \sigma$ for which

$$\frac{|Proj(S, \pi - \sigma)|}{|Proj(S, \pi - \sigma \cup \{i\})|} = UnPred_i(Proj(S, \pi - \sigma)) \leq r.$$

It follows that

$$\frac{|Proj(S, \pi - \sigma \cup \{i\})|}{k^{|\pi - \sigma - \{i\}|}} \geq \frac{1}{r} \cdot \frac{|Proj(S, \pi - \sigma)|}{k^{|\pi - \sigma - \{i\}|}} = \frac{k}{r} \cdot \frac{|Proj(S, \pi - \sigma)|}{k^{|\pi - \sigma|}} \geq \frac{k}{r}$$

Thus the property holds for $\sigma \cup \{i\}$. Eventually, for all $i \in \pi - \sigma$, $UnPred_i(Proj(S, \pi - \sigma)) > r$. Since $Proj(S, \pi - \sigma) \subseteq [k]^{\pi - \sigma}$ and $r < k$, it follows that

$$1 \geq \frac{|Proj(S, \pi - \sigma)|}{k^{|\pi - \sigma|}} \geq \left(\frac{r}{k}\right)^{\ell - |\sigma|}$$

and thus $|\sigma| \leq l$. ■