CS3236 Lecture Notes #1:
Information Measures

Jonathan Scarlett

December 16, 2022

Useful references:

- Cover/Thomas Chapter 2
- MacKay Chapters 2–4

1 Information of an Event

Problem.

- If we are told that random event A occurred (e.g., coin came up tails, two dice added up to 7, it rained today), how much "information" have we learned?
- Approach: Quantify information without any regard to *significance* or *importance*. It is only $\Pr[A]$ that matters.
  - Things like "importance" are usually too subjective to quantify.
- Generically speaking, if A occurs with probability p, then

$$\text{Information}(A) = \psi(p)$$

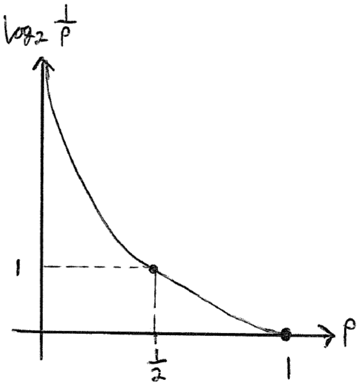
  for some function $\psi(\cdot)$. Perhaps a more intuitive interpretation of $\psi(p)$ is that it quantifies *how surprised we are that event A occurred.* What properties should this function satisfy?

Axiomatic view.

- Here are some very natural properties that we should expect $\psi(p)$ to satisfy:
  1. (Non-negativity) $\psi(p) \ge 0$, i.e., we cannot learn a "negative amount" of information.
  2. (Zero for definite events) $\psi(1) = 0$, i.e., if something was certain to happen, nothing is learned by the fact that it occurred.
  3. (Monotonicity) If $p \le p'$, then $\psi(p) \ge \psi(p')$, i.e., the less likely the event was, the more information is learned by the fact that it occurred.

4. (Continuity) $\psi(p)$ is continuous in p, i.e., small changes in probability don't cause drastic changes in information.
5. (Additivity under independence) $\psi(p_1p_2) = \psi(p_1) + \psi(p_2)$. If A and B are independent events with probabilities p_1 and p_2, then $A \cap B$ has probability p_1p_2, and the information learned from both A and B occurring is the sum of the two individual amounts of information (because they are independent!)

- It can be shown that only $\psi(p) = \log_b \frac{1}{p}$ (for some base $b > 0$) satisfies all three
  - We focus on $b = 2$, which means information is measured in "bits". Another common choice is $b = e$, which means information is measured in "nats".
  - All choices of b are equivalent up to scaling by a universal constant (e.g., number of nats $=$ $(\log_e 2)$ × number of bits). This is much the same as how we can measure distance in meters, kilometers, inches, or miles, but converting from one to another just amounts to scaling.
  - So being told that a probability-p event occurred gives us $\log_2 \frac{1}{p}$ "bits" of information.
  - An illustration:



2 Information of a Random Variable – Entropy

Definition.

- Let X be a discrete random variable with probability mass function (PMF) P_X
- According to the previous section, if we observe $X = x$ then we have learned $\log_2 \frac{1}{P_X(x)}$ bits of information. The **(Shannon) entropy** is simply the average of this value with respect to P_X:

$$\begin{aligned} H(X) &= \mathbb{E}_{X \sim P_X}\left[\log_2 \frac{1}{P_X(X)}\right] \\ &= \sum_x P_X(x) \log_2 \frac{1}{P_X(x)}. \end{aligned}$$

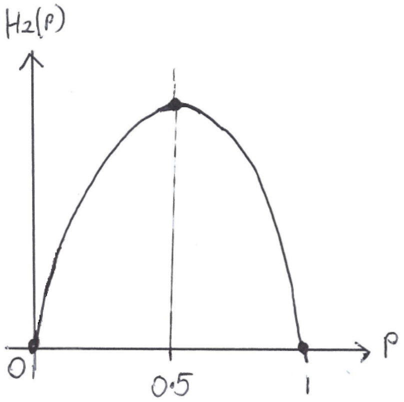
  - Note the convention $0 \log \frac{1}{0} = 0$, which is reasonable since $\lim_{p \to 0} p \log_2 \frac{1}{p} = 0$.

- Can be viewed as a measure of *information in X* or *uncertainty in X* (these are not contradictory)

- **Note.** Here and throughout the vast majority of the course, we only consider *discrete-valued* random variables that can only take on a finite number of values. We will cover *continuous-valued* random variables much later.

Examples.

- Binary source:
  - Suppose $X \sim \text{Bernoulli}(p)$ for some $p \in (0,1)$ (i.e., $P_X(1) = 1 - P_X(0) = p$)
  - Then we get
  $$H(X) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}. \tag{1}$$
  The right hand side, as a function of p, is known as the *binary entropy function.* Since this quantity will be used frequently throughout the course, we give it a formal definition: $H_2(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$ for $p \in [0,1]$. An illustration:



- Uniform source:
  - Suppose X is uniform on a finite set $\mathcal{X}$ (i.e., $P_X(x) = \frac{1}{|\mathcal{X}|}$ for each $x \in \mathcal{X}$, where $|\mathcal{X}|$ is the cardinality of $\mathcal{X}$)
  - Then we get
  $$H(X) = \mathbb{E}\left[\log_2 \frac{1}{1/|\mathcal{X}|}\right] = \log_2 |\mathcal{X}|.$$
  This is intuitive, e.g., with 10 bits we can produce $|\mathcal{X}| = 2^{10}$ combinations of bits.

Axiomatic view [Shannon].

- Suppose that X is a discrete random variable taking N values, with probabilities $\mathbf{p} = (p_1, \dots, p_N)$. If we consider a general information measure of the form
$$\Psi(\mathbf{p}) = \Psi(p_1, \dots, p_N),$$
then what properties should it satisfy?
- Three natural properties:
  1. (Continuity) $\Psi(\mathbf{p})$ is continuous as a function of $\mathbf{p}$. Again, small changes in the distribution don't give large changes in information/uncertainty.

2. (Uniform case) If $p_i = \frac{1}{N}$ for $i = 1, \dots, N$, then $\Psi(\mathbf{p})$ is increasing in N. That is, being uniform over a larger set of outcomes always means more information/uncertainty.
3. (Successive decisions) The following always holds:

$$\Psi(p_1, \dots, p_N) = \Psi(p_1 + p_2, p_3, \dots, p_N) + (p_1 + p_2)\Psi\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right).$$

   This can be viewed as drawing from the distribution on X by first drawing from the corresponding distribution that doesn't distinguish two symbols (the ones with probabilities p_1 and p_2), and then drawing another random variable to resolve those two symbols if needed (which only happens a fraction $p_1 + p_2$ of the time). The total information/uncertainty is the sum of the information/uncertainty from each of the two stages.

- It can be shown that only $\Psi(X) = \text{constant} \times H(X)$ satisfies all three.

Variations.

- Joint entropy of two random variables (X, Y):

$$\begin{aligned} H(X,Y) &= \mathbb{E}_{(X,Y)\sim P_{XY}}\left[\log_2 \frac{1}{P_{XY}(X,Y)}\right] \\ &= \sum_{x,y} P_{XY}(x,y) \log_2 \frac{1}{P_{XY}(x,y)}. \end{aligned}$$

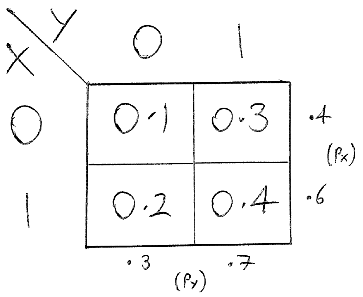
  We can similarly define $H(X,Y,Z)$ or larger collections such as $H(X_1, \dots, X_n)$.

- Conditional entropy of Y given X:

$$\begin{aligned} H(Y|X) &= \mathbb{E}_{(X,Y)\sim P_{XY}}\left[\log_2 \frac{1}{P_{Y|X}(Y|X)}\right] \\ &= \sum_{x,y} P_{XY}(x,y) \log_2 \frac{1}{P_{Y|X}(y|x)} \\ &= \sum_{x} P_X(x) H(Y|X=x), \qquad (2) \end{aligned}$$

  where in the last line, $H(Y|X=x) = \sum_y P_{Y|X}(y|x) \log_2 \frac{1}{P_{Y|X}(y|x)}$ is simply the entropy of the distribution $P_{Y|X}(\cdot|x)$ on Y. We can similarly define quantities like $H(Y_1, Y_2|X_1, X_2)$.

  - Intuition: $H(Y|X=x)$ is the uncertainty in Y after having observed that $X = x$. The conditional entropy $H(Y|X)$ simply averages such a quantity over X, so it represents the average remaining uncertainty in Y after observing X.
  - Example: Consider the joint distributed described as follows.



X \ Y	0	1	(P_X)
0	0.1	0.3	.4
1	0.2	0.4	.6
(P_Y)	.3	.7	

Each entry in the table is a $P_{XY}(x,y)$ value, and the values at the right and bottom are the resulting marginals $P_X(x)$ and $P_Y(y)$ (just add up the relevant row or column).

- Combining the joint and marginal distributions gives the conditionals:

$$P_{Y|X}(0|0) = \frac{P_{XY}(0,0)}{P_X(0)} = \frac{0.1}{0.4} = \frac{1}{4}$$
$$P_{Y|X}(0|1) = \frac{P_{XY}(1,0)}{P_X(1)} = \frac{0.2}{0.6} = \frac{1}{3}.$$

Substitution into the form of $H(Y|X)$ in Eq. (2) gives

$$H(Y|X) = 0.4\,H_2\Big(\frac{1}{4}\Big) + 0.6\,H_2\Big(\frac{1}{3}\Big) \approx 0.8755,$$

where $H_2(p)$ denotes the binary entropy function given on the right-hand side of Eq. (1).

  - Note that $H(Y|X)$ is smaller than $H(Y) = H_2(0.3) \approx 0.8813$ (on average, knowing X reduces uncertainty about Y)
  - But $H(Y|X=1) = H_2\big(\frac{1}{3}\big) \approx 0.9183$ (seeing a *specific* outcome of X may increase uncertainty about Y)

2.1 Properties of Entropy

- **Non-negativity:**

$$H(X) \geq 0$$

with equality if and only if X is deterministic.

  - Intuition: Information/uncertainty cannot be negative
  - Proof: The "information of an event" $\log_2 \frac{1}{p}$ is always non-negative for $p \in [0,1]$, so entropy is the average of a quantity that is always non-negative, and so is itself non-negative. Moreover, only $p = 1$ gives $\log_2 \frac{1}{p} = 0$, so $H(X) = 0$ if and only if X is deterministic.

- **Upper bound:** If X takes values on a finite alphabet $\mathcal{X}$, then

$$H(X) \leq \log_2 |\mathcal{X}|$$

with equality if and only if X is uniform on $\mathcal{X}$. This similarly implies $H(X|Y) \leq \log_2 |\mathcal{X}|$.

  - Intuition: The uniform distribution has the most uncertainty.

- Proof: Let P be the distribution of X, and let Q be the uniform distribution on $\mathcal{X}$, so that $Q(x) = \frac{1}{|\mathcal{X}|}$ for all x. Then note that

$$\begin{aligned}\sum_x P(x) \log_2 \frac{P(x)}{Q(x)} &= \sum_x P(x) \log_2 \big(|\mathcal{X}| \cdot P(x)\big) \\ &= \log_2 |\mathcal{X}| + \sum_x P(x) \log P(x) \\ &= \log_2 |\mathcal{X}| - H(X).\end{aligned}$$

  In Section 3 we will show that the left-hand side is non-negative for *any* distributions P and Q, with equality if and only if $P = Q$. Specialized to the above choices of P and Q, we get $\log_2 |\mathcal{X}| - H(X) \geq 0$ with equality if and only if P is uniform, as desired.

- **Chain rule (two variables):**

$$H(X,Y) = H(X) + H(Y|X)$$

  - Intuition: The overall information in (X,Y) is the information in X plus the remaining information in Y after observing X.
  - Proof: For $(X,Y) \sim P_{XY}$, we have

$$\begin{aligned}H(X,Y) &= \mathbb{E}\bigg[\log \frac{1}{P_{XY}(X,Y)}\bigg] \\ &= \mathbb{E}\bigg[\log \frac{1}{P_X(X)P_{Y|X}(Y|X)}\bigg] \\ &= \mathbb{E}\bigg[\log \frac{1}{P_X(X)} + \log \frac{1}{P_{Y|X}(Y|X)}\bigg] \\ &= H(X) + H(Y|X).\end{aligned}$$

- Chain rule (general):

$$H(X_1, \dots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \dots, X_{i-1}).$$

  - Intuition: Similar to the two-variable case.
  - Proof: Similar to the two-variable case, but instead use the expansion $P_{X_1 \dots X_n} = P_{X_1} \times P_{X_2|X_1} \times P_{X_3|X_1X_2} \times \dots \times P_{X_n|X_1,\dots,X_{n-1}}$.

- Conditioning reduces[1] entropy:

$$H(X|Y) \leq H(X)$$

  with equality if and only if X and Y are independent.

  - Intuition: Having additional information cannot increase uncertainty *on average*.[2]
  - Proof: Equivalent to the property $I(X;Y) \geq 0$ to be proved in Section 4.1.

[1] More precisely, does not increase

[2] In contrast, $H(X|Y=y)$ for a particular y could exceed $H(X)$, as we saw in the example following the conditional entropy definition (note that the roles of X and Y were reversed there).

- Sub-additivity:

$$H(X_1,\dots,X_n) \le \sum_{i=1}^{n} H(X_i)$$

  with equality if and only if $X_1,\dots,X_n$ are independent.

  - Intuition: The uncertainty in several random variables is no more than the sum of individual uncertainty in each one.
  - Proof: Apply "conditioning reduces entropy" to each summand in the general chain rule formula above.

3 A Useful Measure Between Distributions – KL Divergence

- For two PMFs P and Q on a finite alphabet $\mathcal{X}$, the *Kullback-Leibler (KL) divergence* (also known as *relative entropy*) is given by

$$\begin{aligned} D(P\|Q) &= \sum_x P(x)\log_2 \frac{P(x)}{Q(x)} \\ &= \mathbb{E}_{X\sim P}\left[\log_2 \frac{P(X)}{Q(X)}\right]. \end{aligned}$$

- Can be viewed as a kind of "distance" between P and Q, but it is not a distance function in the mathematical sense (in general it is not symmetric and doesn't satisfy the triangle inequality).
- **Claim.** For any distributions P and Q, we have

$$D(P\|Q) \ge 0$$

  with equality if and only if $P = Q$.

  - Proof:

$$\begin{aligned} -D(P\|Q) &= \sum_x P(x)\log \frac{Q(x)}{P(x)} \\ &\overset{(a)}{\le} \sum_x P(x)\left(\frac{Q(x)}{P(x)} - 1\right) \\ &= \sum_x Q(x) - \sum_x P(x) \\ &= 0, \end{aligned}$$

    where (a) uses the inequality $\log\alpha \le \alpha - 1$, which is easily verified graphically. Equality holds in $\log\alpha \le \alpha - 1$ if and only if $\alpha = 1$, which means that equality holds in (a) if and only if $\frac{Q(x)}{P(x)} = 1$ for all x (i.e., $P = Q$).

- The KL divergence (and in fact, also entropy and mutual information) is used extensively in other fields like statistics and machine learning. Some example uses (stated only very roughly here) are:

- In data compression, if the true source is distribution is P but we use an algorithm that wrongly assumes it is Q, then we pay a penalty of $D(P\|Q)$ in the average number of bits per symbol;
- In statistics, if $\mathbf{X} = (X_1, \dots, X_n)$ is i.i.d. with $X_i \sim Q$, then the probability that $\mathbf{X}$ "looks like" it was generated i.i.d. on P (yes, this is quite vague) is roughly $2^{-nD(P\|Q)}$ when n is large. Look up *Sanov's theorem* for a more precise statement.

4 Information Between Random Variables – Mutual Information

Definition.

- Mutual information:

$$I(X;Y) = H(Y) - H(Y|X).$$

- Intuition:
  - $H(Y)$ is the *a priori uncertainty* in Y
  - $H(Y|X)$ is the *remaining uncertainty* in Y after observing X (on average)
  - Hence, $I(X;Y)$ is the *amount of information about Y we learn by observing X* (on average).

Variations.

- Joint version:

$$I(X_1, X_2; Y_1, Y_2) = H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2).$$

- Conditional version:

$$I(X;Y|Z) = H(Y|Z) - H(Y|X, Z).$$

Examples.

1. If X and Y are independent, then it is straightforward to compute $H(Y|X) = H(Y)$, giving $I(X;Y) = 0$ (i.e., independent random variables do not reveal any information about each other).

2. If $Y = X$, then it is straightforward to compute $H(Y|X) = H(X|X) = 0$, and hence $I(X;X) = H(X)$ (i.e., the amount of information a random variable reveals about itself is the entropy).

3. In the example given shortly after Eq. (2), we computed $H(Y|X) \approx 0.8755$ and $H(Y) \approx 0.8813$, which gives $I(X;Y) = H(Y) - H(Y|X) \approx 0.006$.

4. We will see more "insightful" examples when we come to the channel coding (communication) part of the course.

4.1 Properties of Mutual Information

- **Alternative forms:**

$$\begin{aligned}
I(X;Y) &= D(P_{XY}\|P_X \times P_Y) \\
&= \mathbb{E}\bigg[\log_2 \frac{P_{XY}(X,Y)}{P_X(X)P_Y(Y)}\bigg] = \sum_{x,y} P_{XY}(x,y)\log_2 \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \\
&= \mathbb{E}\bigg[\log_2 \frac{P_{Y|X}(Y|X)}{P_Y(Y)}\bigg] = \sum_{x,y} P_{XY}(x,y)\log_2 \frac{P_{Y|X}(y|x)}{P_Y(y)}.
\end{aligned}$$

  - Proof: Substituting $H(Y) = \mathbb{E}\big[\log_2 \frac{1}{P_Y(Y)}\big]$ and $H(Y|X) = \mathbb{E}\big[\log_2 \frac{1}{P_{Y|X}(Y|X)}\big]$ into the definition of mutual information gives $I(X;Y) = \mathbb{E}\big[\log_2 \frac{P_{Y|X}(Y|X)}{P_Y(Y)}\big]$. Multiplying the numerator & denominator by $P_X(X)$ gives $\mathbb{E}\big[\log_2 \frac{P_{XY}(X,Y)}{P_X(X)P_Y(Y)}\big]$, from which the remaining equalities follow easily.

- **Symmetry:** We have

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

and in particular

$$I(X;Y) = I(Y;X)$$

which also implies

$$I(X;Y) = H(X) - H(X|Y).$$

  - Intuition: X and Y reveal an equal amount of information about each other (or maybe this is not that intuitive!)
  - Proof: We have from the above alternative form that

$$\begin{aligned}
I(X;Y) &= \mathbb{E}\bigg[\log_2 \frac{P_{XY}(X,Y)}{P_X(X)P_Y(Y)}\bigg] \\
&= \mathbb{E}\bigg[\log_2 \frac{1}{P_X(X)} + \log_2 \frac{1}{P_Y(Y)} + \log_2 P_{XY}(X,Y)\bigg] \\
&= H(X) + H(Y) - H(X,Y),
\end{aligned}$$

    where we first expanded the logarithm, and then applied the definition of (joint) entropy.

- **Non-negativity:** $I(X;Y) \geq 0$ with equality if and only if X and Y are independent.
  - Intuition: One random variable cannot tell us a "negative amount" of information about the other.
  - Proof: Using the above-established identity $I(X;Y) = D(P_{XY}\|P_X \times P_Y)$, this is just a special case of $D(P\|Q) \geq 0$ with equality if and only if $P = Q$.

- **Upper bounds:** We have

$$\begin{aligned}
I(X;Y) &\leq H(X) \leq \log_2 |\mathcal{X}| \\
I(X;Y) &\leq H(Y) \leq \log_2 |\mathcal{Y}|.
\end{aligned}$$

  - Intuition: The information X reveals about Y (mutual information) is at most the prior information in X (entropy).

- Proof: To show that $I(X;Y) \le H(X)$, combine $I(X;Y) = H(X) - H(X|Y)$ (see above) and $H(X|Y) \ge 0$ (conditional or unconditional entropy is never negative). We already showed $H(X) \le \log_2 |\mathcal{X}|$ earlier, and the remaining claims follow by symmetry, reversing the roles of X and Y.

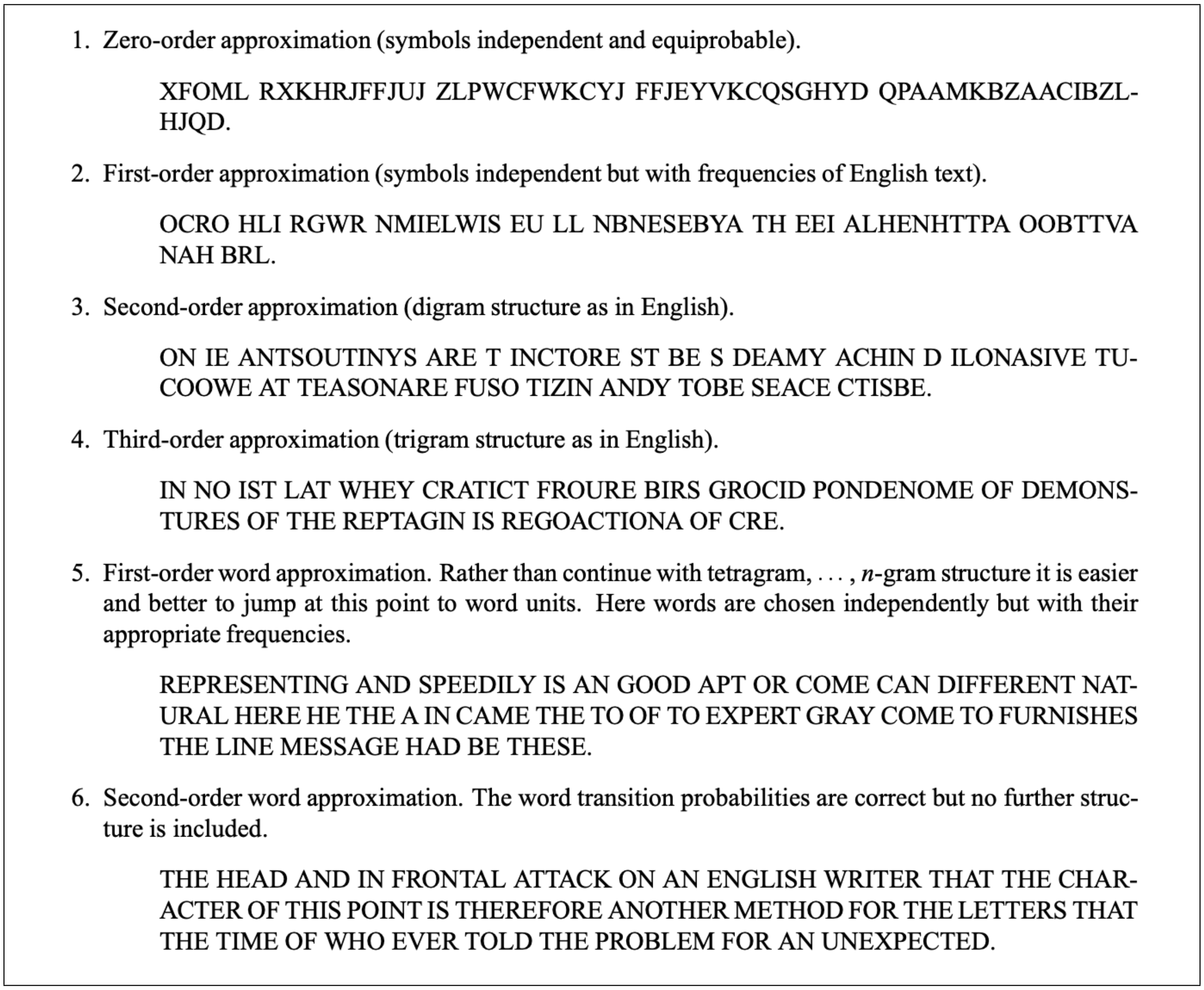
- **Chain rule:**
$$I(X_1, \dots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_1, \dots, X_{i-1}).$$
  - Intuition: Similar to the chain rule for entropy.
  - Proof: Write $I(X_1, \dots, X_n; Y) = H(X_1, \dots, X_n) - H(X_1, \dots, X_n | Y)$ and apply the chain rule for entropy to both terms.

- **Data processing inequality:** If Z depends on (X, Y) only through Y (often stated via the terminology "$X \to Y \to Z$ forms a Markov chain", and equivalent to the statement "X and Z are conditionally independent given Y"), then
$$I(X;Z) \le I(X;Y).$$
  - Intuition: Processing Y (to produce Z) cannot increase the information available regarding X.
  - Proof: As stated above, the statement "Z depends on (X, Y) only through Y" is equivalent to "Z and X are conditionally independent given Y". This means that the property $P_{Z|XY} = P_{Z|Y}$ (as assumed in the result) is equivalent to $P_{X|YZ} = P_{X|Y}$. To deduce the result, we write
$$\begin{aligned} I(X;Z) &\overset{(a)}{=} H(X) - H(X|Z) && (3)\\ &\overset{(b)}{\le} H(X) - H(X|Y,Z) && (4)\\ &\overset{(c)}{=} H(X) - H(X|Y) && (5)\\ &\overset{(d)}{=} I(X;Y), && (6) \end{aligned}$$
    where (a) and (d) use the definition of mutual information, (b) follows since conditioning reduces entropy, and (c) holds because $H(X|Y,Z) = \mathbb{E}\big[\log \frac{1}{P_{X|YZ}(X|Y,Z)}\big] = \mathbb{E}\big[\log \frac{1}{P_{X|Y}(X|Y)}\big] = H(X|Y)$ by the above-established fact $P_{X|YZ} = P_{X|Y}$.
  - Variations: (See the tutorial)
    - If $X \to Y \to Z$ then $I(X;Z) \le I(Y;Z)$.
    - If $W \to X \to Y \to Z$ then $I(W;Z) \le I(X;Y)$.

- **Partial sub-additivity:** If $(Y_1, \dots, Y_n)$ are conditionally independent given $(X_1, \dots, X_n)$, and in addition Y_i depends on $(X_1, \dots, X_n)$ only through X_i, then
$$I(X_1, \dots, X_n; Y_1, \dots, Y_n) \le \sum_{i=1}^{n} I(X_i; Y_i).$$
  However, without the conditional independence assumptions, this property may fail to hold. This will be proved in a later tutorial, and will be important when we get to the topic of channel coding.

5 (Optional) Entropy of English Text

- Shannon's famous 1948 paper discussed several (intentionally over-simplified) probabilistic models for generating English text; see Figure 1 below.



1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONS-TURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Rather than continue with tetragram, . . . , *n*-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NAT-URAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHAR-ACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Figure 1: Excerpt from Shannon's paper.

- Stated differently, #3 generates each letter conditioned on the previous one, #4 conditions on the previous two, #5 lets the "alphabet" $\mathcal{X}$ be the set of all words rather than the set of all characters and generates each word independently, and #6 generates each word conditioned on the previous one.
- **Fundamental question:** How much information (entropy) does each letter of English text tell us?
  - The entropy $H(X)$ of a single character doesn't capture the fact that previous characters help in predicting the next one.
  - As detailed in Chapter 4 of Cover/Thomas, a more meaningful measure in such scenarios is

$$H(X_n|X_1, \dots, X_{n-1}),$$

representing the uncertainty of a given character given all of the previous ones.

- Fitting a model to English text and then calculating the entropy of that model is prone to be inaccurate (too complex to fit a very accurate model!) – is there a simpler approach?

- **Key idea:** The entropy is closely related to *how many guesses are needed (on average) before the correct character is guessed*, by an optimal guessing algorithm.
  - Intuitively, entropy is a measure of uncertainty, and higher uncertainty means more guesses will be needed on average.
  - Writing an optimal guessing algorithm is hard (though an interesting machine learning problem!), so experiments were done under the assumption that *humans are near-optimal guessers*.
  - Using some theory behind the "optimal guessing" viewpoint, and observing the average number of guesses that several humans required, it was estimated that the entropy of English text is only around **1.34 bits per character**
  - Much less than the highest possible value of $\log_2 27 \approx 4.75$ with 27 characters! (*a*-*z* and "space")

- See Chapter 6 of Cover/Thomas for further details.

6 (Optional) Other Properties

More properties of entropy.

- **Functions of random variables:** For a deterministic function f, we have

$$H(f(X)) \le H(X)$$

  - Intuition: Transforming a random variable doesn't increase its information content.
  - Proof: Since $f(X)$ is deterministic given X, we have $H(f(X)|X) = 0$, and hence

$$\begin{aligned} H(X) &= H(X) + H(f(X)|X) \\ &= H(X, f(X)) \\ &= H(f(X)) + H(X|f(X)) \\ &\ge H(f(X)), \end{aligned}$$

    where the first and third lines use the chain rule, and the last line uses non-negativity.
  - An alternative proof is explored in the tutorial.

- **Information-preserving transform:** If Y depends on X only through $f(X)$, then

$$H(Y|X) = H(Y|f(X)).$$

  - Intuition: By the assumption, $f(X)$ already gives us all that X can tell us about Y.
  - Proof: Let $F = f(X)$. By assumption $P_{Y|X}(y|x) = P_{Y|F}(y|f(x))$ for some $P_{Y|F}$, and hence

$$H(Y|X) = \mathbb{E}\left[\log_2 \frac{1}{P_{Y|X}(Y|X)}\right] = \mathbb{E}\left[\log_2 \frac{1}{P_{Y|F}(Y|f(X))}\right] = H(Y|f(X)).$$

- **Concavity:** The entropy $H(X)$ satisfies a useful property known as concavity (as a function of the distribution P_X).

More properties of mutual information.

- **Functions of random variables.** If Y depends on X only through $f(X)$, then

  $$I(X;Y) = I(f(X);Y).$$

  This follows easily from the analogous conditional entropy property above upon applying $I(X;Y) = H(Y) - H(Y|X)$.

- **Convexity properties:** Mutual information $I(X;Y)$ is concave in P_X for fixed $P_{Y|X}$, and is convex in $P_{Y|X}$ for fixed P_X.

  - For an introduction to convexity, see the book "Convex Optimization" by Boyd and Vandenberghe