Both images contain the same fraction of black and white pixels, but the one on the right is far more disordered than the first. We would need far more information to describe the pattern of pixels on the right than on the left. This idea is connected to the signal’s entropy.
Probability and information

Recall we determined the information conveyed by an event occurring is related to its likelihood:

\[
\text{Fundamental Tenet III:} \quad \text{Information content is inversely proportional to probability.}
\]

Today, we will formally define the information \( I_E \) of event \( E \), based on the properties information should have.

This definition of information will also extend naturally to discrete random variables, allowing us to describe information in terms of probability distributions.
Information as a quantity

To begin to quantify information mathematically, what properties should information have?

1) It should be nonnegative: observing something that occurs should not increase uncertainty.

2) The amount of information should decrease as the likelihood increases, as a rare occurrence is more informative than a common one.

3) The amount of information should decrease to zero as the likelihood increases to one, since we already know something certain to occur will occur.

4) If we observe two independent events in succession, the information conveyed by both occurring should be the same as the sum of the information obtained from each event.
   - Recall: what happens to the probabilities of independent events occurring together? What is P(AB)?
Information as a quantity

Proposition: For an event $E$ with probability $P(E)$, we define self-information $I_E$ as

$$I_E = \log_2 \left( \frac{1}{P(E)} \right)$$

Let’s confirm it satisfies the four properties:

1) $I_E$ is nonnegative since $P(E) \leq 1$.
2) $I_E$ is decreasing as $P(E)$ increases.
3) $I_E$ decreases to 0 as $P(E)$ increases to 1.
4) If $A, B$ are independent, then $P(AB) = P(A)P(B)$, and

$$I_{AB} = \log_2 \left( \frac{1}{P(A)P(B)} \right) = \log_2 \left( \frac{1}{P(A)} \right) + \log_2 \left( \frac{1}{P(B)} \right)$$
Information as a quantity

Is this function unique? Essentially, yes. We can change the base of the logarithm, but otherwise, Alfréd Rényi showed that it is the only choice for discrete events that satisfies these properties.

- If the base = 2, we call the corresponding unit of information a “bit”
- If the base = $e$, we call the corresponding unit of information a “nat”

How do we apply to a discrete random variable $X$? Replace $P(E)$ in the formula with the probability mass for the value $x$ associated with the desired outcome.

Rényi is also credited (along with Erdős) with claiming “A mathematician is a device for turning coffee into theorems.”
Proof (optional)

We have \( I(AB) = I(A) + I(B) \). Take a derivative with respect to \( A \):

\[
I(AB) = I(A) + I(B) \Rightarrow \frac{d}{dA} I(AB) = \frac{d}{dA} (I(A) + I(B)) \Rightarrow BI'(AB) = I'(A)
\]

Let \( A = 1 \) to get \( BI'(B) = I'(1) \) for all \( B \), and then integrate:

\[
I(1) - I(x) = \int_x^1 I'(B)\ dB = I'(1) \int_x^1 \frac{1}{B} dB = -I'(1) \ln x
\]

If we additionally constrain \( I(1) = 0 \), then \( I(x) = I'(1) \ln x \).

Since \( I(x) \) is decreasing, \( I'(1) < 0 \) and does not depend on \( x \).

Set \( I'(1) = \frac{1}{\ln b} \) for base \( b > 1 \) to get \( I(x) = \log_b \frac{1}{x} \). QED.
Information examples

Calculate self-information $I_E$ for the following events:

- Draw a king of clubs:

- Roll a pair of dice and get a pair of “1”s:

- Time to play a game...
Information of a binary variable

First, suppose we have a fair coin flip: $P(0) = P(1) = 0.5$. How much information is conveyed by a single coin flip?
- For $X = 0$, we have $I_0 = \log_2(1/0.5) = 1$ bit. For $X = 1$, we have $I_1 = \log_2(1/0.5) = 1$ bit.
- In this case, a single binary coin flip (a bit) gives us 1 bit of information.

Now, suppose the coin were not fair: $P(0) = 0.75$, $P(1) = 0.25$.
- For $X = 0$, we have $I_0 = \log_2(1/0.75) = 0.415$ bits. For $X = 1$, we have $I_1 = \log_2(1/0.25) = 2$ bits.
- So in one case, the coin flip conveys less than 1 bit, and in the other (rarer) case, more than 1 bit.

What about multiple coin flips?
Activity: Guess What (or Who)?

The object of the classic game is to guess your opponent’s secret identity through a series of yes/no questions. We will be playing a simplified game with a collection of 12 food emojis:

Track how few questions you used to narrow down to your opponent’s food item.
Activity: Guess What?

How many questions did you need to ask? What is the average among the class?

Question: How does this compare versus $\log_2(12) = 3.585$?
Average information and entropy

Computing the average information of a random variable is as simple as combining the information for each possible outcome, weighted by the likelihood of that outcome:

\[ E\{I_X\} = \sum_x I_X(x) \cdot P(X = x) \]

This average information is also called the entropy \( H(X) \) of the random variable. This term arises from statistical physics:

“My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage.'”

Average information and entropy

Let’s go back to the coin flip example.

- Fair coin: \( I_0 = I_1 = 1 \) bit, so \( H = 1 \) bit as well.
- For biased coin: \( I_0 = 0.415 \) bits, \( I_1 = 2 \) bits, so \( H = (0.75)(0.415) + (0.25)(2) = 0.811 \) bits. On average, a biased coin conveys less information than a fair coin!
- Why is this intuitive?

How about a fair six-sided die?

- \( I_1 = I_2 = I_3 = I_4 = I_5 = I_6 = \log_2(6) = 2.585 \) bits
- The entropy is the average = 2.585 bits.

What about drawing a playing card?
Average information and entropy

Minimum entropy value is zero. How can this be attained?

- One outcome is certain: on average, no information is gained by observing it.

Maximum entropy value occurs when all values are equally likely (we call this a uniform distribution).

- For a fair coin, this is 1 bit.
- For a fair six-sided die, this is \( \log_2(6) = 2.585 \) bits.
- For \( N \) possible equally likely outcomes, \( H = \log_2(N) \) bits.

However, realistic signals rarely have either absolute certainty or uniform distributions. So their entropy is somewhere in between. This is good news!
Entropy of real signals

Let’s examine a “standard” image: lots of details

Zoom in closer: lots of redundancy!

Data compression will take advantage of this redundancy to reduce storage/transmission requirements.
Entropy of real signals

Let’s simplify things to a black/white example:

Total pixels: 31 x 24 = 744 pixels

We could store each pixel as a separate bit, totaling 744 bits of storage. This would be consistent with equiprobable 0’s and 1’s.

Instead, let’s ask how much information is stored in this image: 571 black, 173 white pixels.
Entropy of real signals

For each outcome (B or W), count the relative frequencies: \( P(B) = \frac{571}{744} = 0.767 \), \( P(W) = \frac{173}{744} = 0.233 \).

For each outcome, also compute self-information: \( I_B = 0.382 \) bits, \( I_W = 2.104 \) bits.

Average to find entropy or average information: \( H = 0.783 \) bits per pixel.

Multiply by \# of pixels to find total information: \((0.783 \text{ bits/pixel}) \times (744 \text{ pixels}) = 583 \) bits.

Two interpretations of entropy:

- **Abstract**: average information among outcomes according to a probability distribution.
- **Empirical**: average information among realizations (e.g., pixels) according to relative frequencies.

Question for next time: how do we code or compress data to minimize redundancy?
Sharing information

In communications and storage, we want the signal we transmit or write to be the same as the signal received or read from the medium.

In general, we cannot expect an error-free representation, since real signals get corrupted by noise.

One way of quantifying information loss is to measure how much information is shared between the transmitted and received signal. This measure of shared information is called *mutual information*.
Mutual information

To compute mutual information, first we define conditional entropy, $H(X|Y)$ that represents the average information remaining to be gained from $X$ after observing a related signal $Y$.

First, define a conditional expected value $E_{X|Y}\{f(X) \mid Y = y\}$, which is a function of random variable $Y$, and thus is itself random.

$$E_{X|Y}\{f(X)\mid Y = y\} = \sum_x f(x) \cdot P(X = x|Y = y)$$

The conditional entropy $H(X|Y)$ is the expected value (with respect to the remaining random variable $Y$) of this random function, with $f(x) = \log_2 \frac{1}{P(X = x|Y = y)}$

$$H(X|Y) = E \left\{ E_{X|Y} \left\{ \log_2 \frac{1}{P(X = x|Y = y)} \right\} \right\} = \sum_y P(Y = y) \sum_x P(X = x|Y = y) \cdot \log_2 \frac{1}{P(X = x|Y = y)}$$
Mutual information

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

This is the difference between the information available from observing just $X$, and the remaining information in $X$ after observing $Y$.

Mutual information is symmetric, but conditional entropy is not. $H(X|Y)$ is not equal to $H(Y|X)$, but $I(X;Y) = I(Y;X) = H(Y) - H(Y|X)$.

Let’s try an example: $Y = X \oplus E$, where $X$ and $E$ are independent binary random variables, and $\oplus$ is the binary “exclusive or” operator. We’ll go over the logical interpretation of this operator later in the semester, but here we just think of it mathematically:

$$Y = \begin{cases} X, & E = 0 \\ 1 - X, & E = 1 \end{cases}$$

Let’s try two cases: $P(E = 1) = 0.5$, $P(E = 1) = 0.1$. Assume $P(X = 0) = P(X = 1) = 0.5$. 


Mutual information (example)

First, \( P(E = 1) = 0.5 \). What is \( P(Y|X) \)? What is \( P(Y) \)?

Since \( Y \) depends on both \( X \) through \( E \), which has \( P(E=1) = 0.5 \).

\[
P(Y=0|X=0) = P(E=0) = 0.5, \text{ and } P(Y=1|X=0) = P(E=1) = 0.5.
P(Y=0|X=1) = P(E=1) = 0.5, \text{ and } P(Y=1|X=1) = P(E=0) = 0.5.
\]

To find \( P(Y) \), we use
\[
P(Y) = P(Y=0) = P(X=0) P(Y=0|X=0) + P(X=1) P(Y=0|X=1) = 0.5*0.5 + 0.5*0.5 = 0.5
\]

What is \( H(Y|X) \)? What is \( H(Y) \)? What is \( I(X;Y) \)?

\[
H(Y|X) = P(X=0) * \left[ P( Y=0|X=0 ) * \log_2 \left( \frac{1}{P(Y=0|X=0)} \right) + P( Y=1|X=0 ) * \log_2 \left( \frac{1}{P(Y=1|X=0)} \right) \right] + \\
P(X=1) * \left[ P( Y=0|X=1 ) * \log_2 \left( \frac{1}{P(Y=0|X=1)} \right) + P( Y=1|X=1 ) * \log_2 \left( \frac{1}{P(Y=1|X=1)} \right) \right] = 1 \text{ bit}
\]

\[
H(Y) = P(Y=0) \log_2 \left( \frac{1}{P(Y=0)} \right) + P(Y=1) \log_2 \left( \frac{1}{P(Y=1)} \right) = 1 \text{ bit},
\]

so \( I(X;Y) = 0 \text{ bit (i.e., } X \text{ and } Y \text{ share no information because } X = Y \text{ is just as likely as } X \text{ not equal to } Y \)
Mutual information (example)

Now, $P(E = 1) = 0.1$. What is $P(Y|X)$? What is $P(Y)$?

Since $Y$ depends on both $X$ through $E$, which has $P(E=1) = 0.1$.

$P(Y=0|X=0) = P(E=0) = 0.9$, and $P(Y=1|X=0) = P(E=1) = 0.1$.

$P(Y=0|X=1) = P(E=1) = 0.1$, and $P(Y=1|X=1) = P(E=0) = 0.9$.

To find $P(Y)$, we use $P(Y=0) = P(X=0) P(Y=0|X=0) + P(X=1) P(Y=0|X=1) = 0.5*0.9 + 0.5*0.1 = 0.5$ (same as before)

What is $H(Y|X)$? What is $H(Y)$? What is $I(X;Y)$ now?

$$H(Y|X) = P(X=0) \cdot \left[ P(Y=0|X=0) \cdot \log_2 \left( \frac{1}{P(Y=0|X=0)} \right) + P(Y=1|X=0) \cdot \log_2 \left( \frac{1}{P(Y=1|X=0)} \right) \right] +$$
$$P(X=1) \cdot \left[ P(Y=0|X=1) \cdot \log_2 \left( \frac{1}{P(Y=0|X=1)} \right) + P(Y=1|X=1) \cdot \log_2 \left( \frac{1}{P(Y=1|X=1)} \right) \right] = 0.469 \text{ bits}$$

$$H(Y) = P(Y=0) \log_2 \left( \frac{1}{P(Y=0)} \right) + P(Y=1) \log_2 \left( \frac{1}{P(Y=1)} \right) = 1 \text{ bit (same as before),}$$

so $I(X;Y) = 0.531 \text{ bits (i.e., X and Y share some information)}$
Mutual information and capacity

In the example, we’ve assumed our source signal $X$ is a simple fair binary variable, which we’ve seen requires a lot of information to encode. A natural question is how much information can be communicated from transmitter to receiver, in general. Assuming we can optimize $P(X)$ for our communication system, we define the capacity of the system to communicate information as the maximum mutual information as a function of $P(X)$:

**Fundamental Tenet IV:**

*The capacity of a communications channel (or system) is the maximum mutual information between source and receiver that can be communicated with respect to the source distribution $P(X)$.***

We will see other ways to measure capacity later in the semester.
More about capacity

This definition motivates a number of questions we will address later in the course:

◦ How do we design P(X)?
◦ How do we characterize a communications (or storage) system this way?
◦ Is this an absolute limit? Or can we exceed it?
◦ How close can we get to capacity in practice?

The key innovation in mathematical communications theory was not only the creation of this limit, but the strategies that followed to achieve it!

Image credits: (left) Marie Gallager; (right) Télécom Bretagne.

Robert Gallager and Claude Berrou developed low density parity check codes and turbo codes, respectively, which provide performance close to information-theoretic capacity.
Announcements

Next time: Compression and coding methods

Homework #5 will go out next week.