

Lecture 1: Probability and Entropy

ENGR 76 lecture notes — March 31, 2026

David Tse, Stanford University

Overview

These notes combine the first-half slide presentation with the second-half lecture. The goal is to connect the motivating applications — especially image compression and communication — to the mathematical language of *random variables*, *probability mass functions*, *expectation*, and *entropy*.

1 Why information science?

Before information theory, communication systems were designed in a source-specific and medium-specific way. Telegraphy, telephony, radio, and television were all developed as separate technologies. Claude Shannon's 1948 insight was that both the *source* of information and the *channel* carrying that information can be modeled probabilistically.

Once that abstraction is made, the same mathematical framework applies to many different settings:

- **Storage and compression:** How many bits are really needed to represent an image, audio clip, or video?
- **Communication:** At what rate can bits be transmitted reliably over a noisy medium?

This leads to three central ideas highlighted in the slides:

1. The minimum compression rate of a source is governed by its **entropy**.
2. The maximum reliable communication rate of a channel is its **capacity**.
3. End-to-end reliable communication is possible when source entropy is below channel capacity.

2 Motivating example: image compression

The slides used digital images to motivate the question “What is information?” This is a natural place to connect the first half of the lecture to the second.

2.1 Raw representation

A grayscale pixel intensity can be modeled as a variable taking values in

$$\{0, 1, 2, \dots, 255\},$$

so one raw grayscale pixel needs 8 bits. A color image has three channels (red, green, blue), so a raw RGB pixel uses

$$3 \times 8 = 24 \text{ bits} = 3 \text{ bytes}.$$

For a typical image of size 4000×3000 :

$$\text{number of pixels} = 4000 \cdot 3000 = 12,000,000,$$

so the raw storage is

$$12,000,000 \times 3 \text{ bytes} = 36,000,000 \text{ bytes} \approx 36 \text{ MB}.$$

A one-hour raw video at 1920×1080 , 30 frames/sec, RGB, 8 bits/channel is even larger, on the order of hundreds of GB. So a central question is:

Can the same information be represented using fewer bits?

To answer that, we need a mathematical model for uncertainty and information.

3 Random variables

A **random variable** X is a variable whose value is uncertain, but which takes values in a specified set.

In this lecture, the examples were all discrete, so X takes values in a finite or countable set.

3.1 Examples from the lecture

Example 1. Binary source. A bit can be modeled as

$$X \in \{0, 1\},$$

with

$$P(X = 0) = p, \quad P(X = 1) = 1 - p.$$

Example 2. Grayscale pixel intensity. A pixel in a grayscale image can be modeled as

$$X \in \{0, 1, 2, \dots, 255\}.$$

Here X is the intensity value of the pixel.

The point is that information science begins by abstracting a physical object into a mathematical variable with possible outcomes.

4 Probability mass function (pmf)

Once the set of possible values is specified, the next step is to describe how likely each value is.

Definition

For a discrete random variable X , the **probability mass function** (pmf) is

$$p_X(i) = P(X = i).$$

It satisfies

$$p_X(i) \geq 0, \quad \sum_i p_X(i) = 1.$$

4.1 Uniform pmf

A special case is the **uniform pmf**. If X can take m values and all are equally likely, then

$$P(X = i) = \frac{1}{m}, \quad i = 0, 1, \dots, m - 1.$$

This corresponds to maximum uncertainty among all distributions on m symbols.

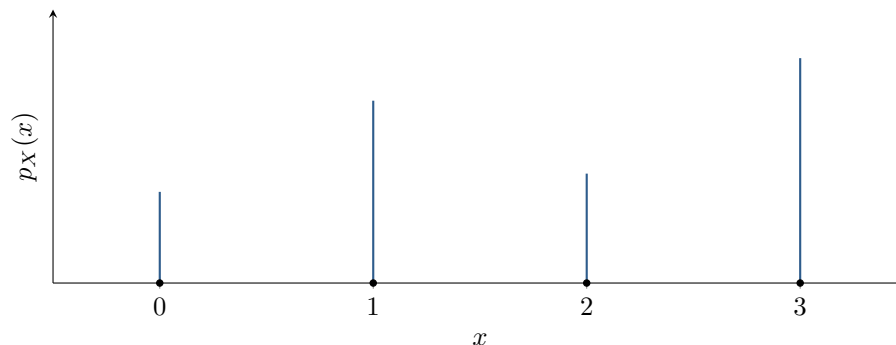


Figure 1: A schematic pmf: a discrete random variable assigns probability mass to each possible value.

4.2 Examples

Binary source. If $X \in \{0, 1\}$ and $P(X = 0) = p$, then the pmf is fully specified by the single number p .

Pixel intensity. If $X \in \{0, 1, \dots, 255\}$ represents a grayscale pixel, then the pmf tells us how often each intensity occurs. For a dark image, the pmf may place more weight on small intensities; for a bright image, more weight on large intensities; for a highly structured image, the pmf is usually far from uniform.

5 Events

An **event** is a set of outcomes.

If X is a random variable, then statements such as

$$X = 1$$

and

$$X \in \{0, 1, \dots, 127\}$$

are events.

The second event is directly related to the image example: if X is pixel intensity, then

$$X \in \{0, 1, \dots, 127\}$$

means “the pixel lies in the darker half of the grayscale range.”

Events are how we ask questions about the random variables. For example, What is the probability the pixel is dark? $P(X \leq 127)$.

6 Expectation

The **expectation** of a real-valued discrete random variable X is its probability-weighted average:

$$E[X] = \sum_i i P(X = i).$$

If $X \in \{0, 1, \dots, m-1\}$, this becomes

$$E[X] = \sum_{i=0}^{m-1} i P(X = i).$$

6.1 Interpretation

Expectation is not necessarily a value that X itself takes; rather, it is the average value of X under repeated sampling.

Example: binary source. If $X \in \{0, 1\}$ with $P(X = 0) = p$ and $P(X = 1) = 1 - p$, then

$$E[X] = 0 \cdot p + 1 \cdot (1 - p) = 1 - p.$$

So the expected value equals the probability that the bit is 1.

6.2 Functions of a random variable

A very important extension is: if

$$Y = f(X),$$

then

$$E[Y] = E[f(X)] = \sum_i f(i) P(X = i).$$

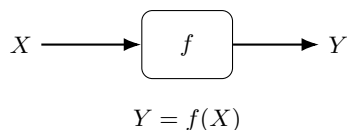


Figure 2: A function of a random variable: first sample X , then apply a deterministic map f to obtain Y .

Example: thresholding a pixel. Let $X \in \{0, 1, \dots, 255\}$ be a grayscale pixel and define

$$f(x) = \begin{cases} 1, & x \geq 128, \\ 0, & x < 128. \end{cases}$$

Then $Y = f(X)$ is a binary variable indicating whether the pixel is bright. Its expectation is

$$E[Y] = P(X \geq 128),$$

which is exactly the probability that the pixel lies in the bright half of the intensity range.

7 Entropy

Expectation summarizes the *average value* of a random variable. Entropy summarizes its *average uncertainty*. In information theory, entropy is a central quantity because it measures the amount of uncertainty in a source.

Definition

If X is a discrete random variable with pmf p_X , then its **entropy** is

$$H(X) = \sum_i p_X(i) \log_2 \frac{1}{p_X(i)} = - \sum_i p_X(i) \log_2 p_X(i).$$

The unit is **bits** when the logarithm is base 2.

We have now introduced the definition of entropy. In future lectures, we will show that entropy gives the fundamental limit of lossless compression.

7.1 Binary entropy

For a binary random variable $X \in \{0, 1\}$ with

$$P(X = 0) = p, \quad P(X = 1) = 1 - p,$$

the entropy is

$$H(p) = p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{1 - p}.$$

This is the binary entropy function.

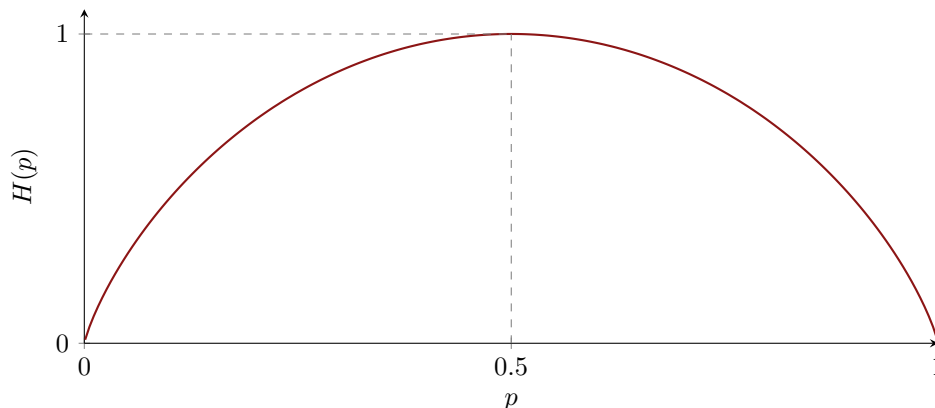


Figure 3: The binary entropy function. It is 0 at $p = 0$ and $p = 1$, and it is maximized at $p = 1/2$, where $H(p) = 1$ bit.

7.2 Interpretation of the curve

Three qualitative facts are worth emphasizing:

- If $p = 0$ or $p = 1$, the outcome is certain, so the entropy is 0.
- If $p = 1/2$, the two outcomes are equally likely, so the uncertainty is largest and the entropy is 1 bit.
- Entropy depends only on the probability distribution, not on the physical meaning of the symbols.

8 Connecting entropy back to compression

Why should entropy matter for compression?

Suppose a source produces symbols according to a pmf p_X . If some symbols are much more likely than others, then the source output is *structured*; it contains less uncertainty than a uniform source on the same alphabet. This suggests that one should be able to represent typical outputs using fewer bits than a naive fixed-length description.

8.1 Example: grayscale pixels

A raw grayscale pixel always uses 8 bits because there are 256 possible values. But if the pixel distribution is highly nonuniform, then the *average information content per pixel* may be much smaller than 8 bits. This is why statistical compression is possible.

8.2 Example: images

A raw 4000×3000 RGB image uses about 36 MB, but real images have massive structure:

- neighboring pixels are correlated,
- natural scenes do not use all RGB combinations equally often,
- the eye is more sensitive to some distortions than others.

These facts are exactly what image compression algorithms exploit. Later in the course this will lead to lossless compression, transform coding, quantization, and practical formats.

9 Summary

The first lecture sets up the core language of the course.

1. Information science begins by modeling sources probabilistically.
2. A **random variable** describes the possible outputs of a source.
3. A **pmf** specifies how likely each output is.
4. An **event** is a set of outcomes of interest.
5. **Expectation** measures average numerical value.
6. **Entropy** measures average uncertainty, in bits.
7. In future lectures, we will show that entropy determines the fundamental limit of lossless compression.

Takeaway

The image-compression examples from the slides are not just applications; they motivate why probability and entropy matter. Once a source is modeled statistically, the question “How much can I compress it?” becomes a precise mathematical question.