INFORMATION THEORY AND APPLICATIONS IN COMPUTER SCIENCE

Lecture 1: Entropy and Information

Draft · Sections 1.1–1.5

16 September 2026

Conventions. Random variables and channel alphabets are finite throughout. All logarithms are base two, except ln; all information quantities are in bits. We set $0 \log 0 = 0$, omit zero-probability outcomes from expectations, and define conditional distributions only at events of positive probability. Write $X^n = (X_1, \dots, X_n)$, $X_{<i} = (X_1, \dots, X_{i-1})$, and $[M] = \{1, \dots, M\}$. The binary entropy is $h_2(t) = -t \log t - (1-t) \log(1-t)$, with endpoint values zero.

1 Entropy, information, and lossless coding

How much uncertainty is there in a random choice? How much does an observation resolve? How many bits are needed to describe the outcome? This lecture develops one quantity that connects these questions.

Shannon's 1948 paper *A Mathematical Theory of Communication* made the statistical structure of a message source central to communication. A source chooses a message; a receiver must reproduce it. More predictable sources should require fewer bits, even when their alphabets are equally large. The mathematical problem concerns the choices a source can make and their probabilities, rather than the semantic importance of a particular message.[1]

1.1 Entropy as expected surprise

An outcome of probability $1/2$ is less surprising than one of probability $1/8$. To quantify this, assign a surprise $s(p)$ to an event of probability $p > 0$. We want independent events to have additive surprises:

$$s(pq) = s(p) + s(q).$$

We also want certainty to have surprise zero, rarer events to be more surprising, and surprise to vary continuously with probability. Choose the unit by setting $s(1/2) = 1$. These requirements force

$$\boxed{s(p) = \log \frac{1}{p}.}$$

Here is a short derivation. The function $f(t) = s(2^{-t})$, for $t \geq 0$, is continuous, additive, and satisfies $f(1) = 1$. Additivity gives $f(a/b) = a/b$ for nonnegative integers a and positive integers b. Continuity extends this to every real $t \geq 0$, so $f(t) = t$. A zero-probability outcome has no finite surprise, but it never occurs and will not contribute to the expectation.

Definition 1.1 (Entropy)**.** For a finite random variable X with distribution p, its *entropy* is its expected surprise:

$$H(X) = H(p) = \mathbb{E} \log \frac{1}{p(X)} = - \sum_{x : p(x) > 0} p(x) \log p(x).$$

[1] Claude E. Shannon, *A Mathematical Theory of Communication* (1948), especially the introduction and Section 6.

For a uniform choice among m possibilities, each outcome has surprise $\log m$, so $H(X) = \log m$. This is the number of binary choices needed when m is a power of two. For nonuniform choices, entropy averages the different surprises; it need not be an integer.

Consider four symbols with probabilities $(1/2, 1/4, 1/8, 1/8)$. Their surprises are $1, 2, 3, 3$ bits, and

$$H(X) = \tfrac{1}{2} \cdot 1 + \tfrac{1}{4} \cdot 2 + \tfrac{1}{8} \cdot 3 + \tfrac{1}{8} \cdot 3 = \tfrac{7}{4}.$$

Two bits always suffice to identify a symbol. The codewords $0, 10, 110, 111$ instead use $7/4$ bits on average. They can be concatenated and decoded without punctuation. Later we will prove that no prefix code has a smaller expected length. For now, the coincidence between surprise and code length is a reason to investigate entropy.

An alternative starting point: uncertainty in stages. One can characterize the entropy of a whole distribution without first assigning surprise to individual outcomes. If a choice is made by first selecting group i with probability p_i, then selecting within that group according to q_i, uncertainty should obey the *grouping rule*

$$U\big((p_i q_{ij})_{i,j}\big) = U(p) + \sum_i p_i U(q_i).$$

For example, distinguish the first symbol of our four-symbol source from the remaining three. The group choice is a fair bit. Within the second group the probabilities are $(1/2, 1/4, 1/4)$, whose entropy is $3/2$. Thus total uncertainty is $1 + (1/2)(3/2) = 7/4$. Grouping, continuity, symmetry, and a normalization of uniform choices determine $U = H$. We will soon recognize grouping as the entropy chain rule.

1.2 Basic properties: uncertainty, support, and randomness

Proposition 1.2 (Range and equality cases)**.** *If $S = \operatorname{supp} p$, then*

$$0 \le H(X) \le \log |S|. \tag{1.1}$$

Entropy is zero exactly when X is deterministic. The upper bound is attained exactly when X is uniform on its support.

Proof. Each summand $p(x)\log(1/p(x))$ is nonnegative and is strictly positive when $0 < p(x) < 1$. Their sum is zero exactly when one outcome has probability one. For the upper bound, concavity of log gives

$$H(X) = \mathbb{E} \log \frac{1}{p(X)} \le \log \mathbb{E} \frac{1}{p(X)} = \log \sum_{x \in S} 1 = \log |S|.$$

Strict concavity makes equality equivalent to $1/p(X)$ being constant on S, which means uniformity. □

For a bit equal to one with probability t,

$$H(X) = h_2(t) = -t \log t - (1-t)\log(1-t).$$

Figure 1 shows the whole range: certainty at either endpoint, and maximum uncertainty at a fair bit. Interchanging the labels zero and one changes no uncertainty.

Entropy is often called a measure of randomness. This refers to a *distribution*, not to the visual appearance of an individual string. A fixed, known string has entropy zero, however irregular it looks. A uniform random n-bit string has entropy n, including the possibility of drawing an orderly-looking string. A model with temporal dependence can also have fair one-bit marginals

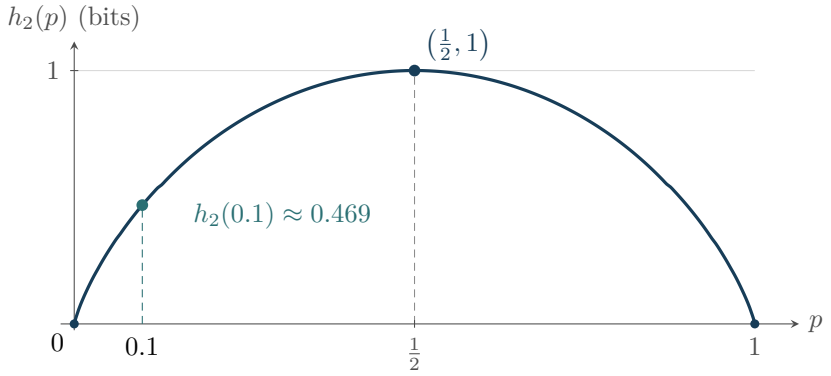

Figure 1: Binary entropy is zero for a deterministic bit and reaches its maximum of one bit at $p = 1/2$. The symmetry $h_2(p) = h_2(1-p)$ reflects interchanging the two bit labels.

and little joint entropy: if every coordinate repeats the same fair bit, the entire string has entropy one. Entropy depends on what is random and what information is already available.

There is a precise connection to statistical mechanics: for a specified finite distribution over physical microstates, Gibbs entropy is Shannon entropy multiplied by $k_B \ln 2$.

1.3 Joint entropy, conditioning, and the chain rule

The pair (X, Y) is itself a random variable. Its *joint entropy* is

$$H(X,Y) = -\sum_{x,y:p(x,y)>0} p(x,y) \log p(x,y).$$

An observation $Y = y$ changes the distribution of X to $p_{X|Y=y}$. The *conditional entropy* averages the uncertainty remaining after this observation:

$$H(X \mid Y) = \sum_{y:p_Y(y)>0} p_Y(y) H(p_{X|Y=y}) = \mathbb{E} \log \frac{1}{p(X \mid Y)}.$$

This is an average over Y, not the entropy at one particular outcome.

Factoring $p(x,y) = p_X(x)p(y \mid x)$ on the joint support gives $-\log p(x,y) = -\log p_X(x) - \log p(y \mid x)$. Taking expectations, and then exchanging X, Y, proves the *chain rule*:

$$\boxed{H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y).} \tag{1.2}$$

Describe the first outcome, then describe what remains uncertain about the second. The total does not depend on which is described first.

Repeating this factorization, also within each conditional distribution, gives

$$H(X^n \mid Z) = \sum_{i=1}^{n} H(X_i \mid Z, X_{<i}).$$

Without Z, this is the ordinary entropy chain rule. If the coordinates are independent, each conditional distribution equals its marginal distribution, so $H(X^n) = \sum_i H(X_i)$. In particular, n independent fair bits have entropy n, whereas $H(X,X) = H(X)$.

Proposition 1.3 (Deterministic processing cannot increase entropy)**.** *If $Y = f(X)$, then $H(Y) \le H(X)$. Equality holds exactly when f is one-to-one on the support of X.*

Proof. Since $H(Y \mid X) = 0$, the chain rule gives

$$H(X) = H(X, Y) = H(Y) + H(X \mid Y) \ge H(Y).$$

Equality requires every positive-probability conditional distribution of X given $Y = y$ to be a point mass. This is exactly the stated injectivity condition. □

A deterministic function can merge possible outcomes, but it cannot create new uncertainty. Randomized processing can: a constant input may be mapped to an independent fair bit. The information about an earlier source still cannot increase, as we will prove below.

1.4 Conditioning reduces entropy on average

Proposition 1.4 (Concavity of entropy)**.** *For distributions $p^{(j)}$ on a common finite alphabet and weights $\lambda_j \ge 0$ summing to one,*

$$H\left(\sum_j \lambda_j p^{(j)}\right) \ge \sum_j \lambda_j H(p^{(j)}).$$

Equality holds exactly when all distributions having positive weight agree.

Proof. The function $\phi(t) = -t \log t$, extended continuously by $\phi(0) = 0$, is strictly concave on $[0, 1]$: in the interior, $\phi''(t) = -1/(t \ln 2) < 0$. Apply concavity coordinate by coordinate and sum. For equality, each coordinate must agree across all positive-weight components. □

The marginal is the mixture of its conditional distributions: $p_X = \sum_y p_Y(y) p_{X|Y=y}$. Applying concavity proves

$$\boxed{H(X \mid Y) \le H(X),}$$

with equality exactly when every positive-probability observation leaves the distribution of X unchanged, or equivalently when X, Y are independent. Applying the same argument after conditioning on each $Z = z$ gives

$$H(X \mid Y, Z) \le H(X \mid Z).$$

Combining this with the chain rule gives *subadditivity*: $H(X, Y) \le H(X) + H(Y)$, with equality exactly for independence.

The averaging matters. Suppose $\Pr[X = 0] = 0.9$ and $\Pr[X = 1] = \Pr[X = 2] = 0.05$, and observe $Y = \mathbf{1}\{X \ne 0\}$. Initially $H(X) = h_2(0.1) + 0.1 \approx 0.569$. If $Y = 1$, the remaining possibilities are equiprobable, so $H(X \mid Y = 1) = 1$: this particular observation increases uncertainty. But $H(X \mid Y) = 0.9 \cdot 0 + 0.1 \cdot 1 = 0.1$, which is smaller than $H(X)$.

1.5 Mutual information: uncertainty resolved

Definition 1.5. The *mutual information* of X, Y is

$$I(X; Y) = H(X) - H(X \mid Y).$$

Their *conditional mutual information* given Z is

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

The chain rule and the conditioning inequality immediately imply

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y \mid X),$$
$$0 \le I(X;Y) \le \min\{H(X), H(Y)\}.$$

Thus information is symmetric, even though we introduced it as the effect of observing Y on uncertainty about X. It vanishes exactly when the two variables are independent. An observation with at most 2^b possible values reveals at most b bits.

The conditional version is likewise symmetric and nonnegative. It is the average mutual information in the distributions conditioned on $Z = z$, and it vanishes exactly when X, Y are independent in every such distribution of positive probability.

Example 1.6 (Shared independent bits)**.** Let $Z_1, \dots, Z_7$ be independent fair bits, and set $X = (Z_1, \dots, Z_5)$, $Y = (Z_4, \dots, Z_7)$. Then $H(X) = 5$, $H(Y) = 4$, and $H(X,Y) = 7$, because the pair determines exactly the seven independent bits. Consequently

$$I(X;Y) = 2, \qquad H(X \mid Y) = 3, \qquad H(Y \mid X) = 2.$$

The common bits Z_4, Z_5 explain the overlap literally in this example. For general variables the same entropy identities hold, without requiring a representation as shared independent bits.

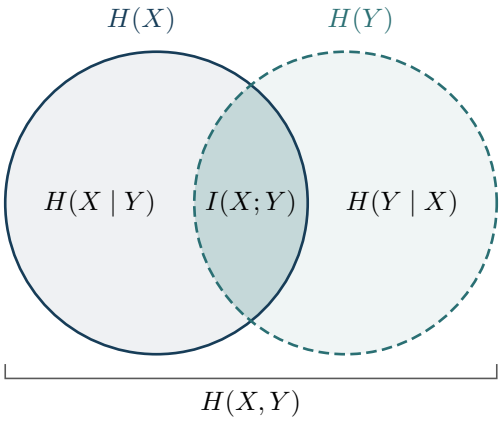

Figure 2: A mnemonic for the entropy identities of two random variables. The regions represent information quantities; their areas are schematic. They are not probability events, and the picture does not assert a general Venn representation for three or more variables.

Information also has a chain rule. Insert and subtract $H(X \mid Y)$ in the entropy difference defining $I(X;Y,Z)$. This gives

$$I(X;Y,Z) = I(X;Y) + I(X;Z \mid Y), \tag{1.3}$$
$$I(X^n;Y) = \sum_{i=1}^{n} I(X_i;Y \mid X_{<i}).$$

The second identity follows by subtracting the conditional entropy chain rule given Y from the unconditional one. No independence is needed. An additional observation contributes the information it carries *conditional on what is already known.*

Example 1.7 (Information can be present only jointly)**.** Let A, B be independent uniform bits and let $Z = A \oplus B$. For either value of A, Z is uniform, so $I(A; Z) = 0$; similarly $I(B; Z) = 0$. But Z is determined by (A, B), so $I(A, B; Z) = H(Z) = 1$. The chain rule reads

$$1 = I(A; Z) + I(B; Z \mid A) = 0 + 1.$$

Once A is known, observing Z determines B. Pairwise independence does not imply joint independence, and conditioning can increase mutual information even though it reduces entropy on average. This is one reason not to interpret a three-variable information diagram as an ordinary Venn diagram with necessarily nonnegative regions.

Theorem 1.8 (Data processing for information)**.** *Suppose Z is produced from Y without further access to X: the joint distribution factors as $p(x, y)K(z \mid y)$, where $K(\cdot \mid y)$ is a probability distribution. Then*

$$I(X; Z) \leq I(X; Y).$$

Proof. The factorization says that X, Z are conditionally independent given Y, so $I(X; Z \mid Y) = 0$. Expanding in two orders gives

$$I(X; Y) = I(X; Y, Z) = I(X; Z) + I(X; Y \mid Z) \geq I(X; Z).$$

□

We write this condition as the Markov chain $X \to Y \to Z$. It includes computing a statistic of Y or adding fresh independent noise. The inequality limits information about X, not the entropy of the output Z. Equality means $I(X; Y \mid Z) = 0$: after seeing Z, the discarded observation Y offers no further information about X.