Entropy in Discrete and Continuous Spaces and a Relation to the Heat Equation

Tran Duc Huy

Abstract

This paper is a self-contained introduction to entropy, from Shannon's discrete measure of uncertainty to the continuous theory and its connection with the heat equation. The discrete part covers the definition, its basic properties, a theorem of Khinchin that identifies it uniquely, and the related quantities of relative entropy and mutual information. On the continuous line, differential entropy, the direct analogue of Shannon's, loses two properties the discrete version has. It can be negative, and it changes under a change of coordinate. Relative entropy keeps both and is the quantity that extends cleanly. The endpoint is a relation between entropy and the heat equation. The Gaussian supplies it, as both the density of greatest entropy at a fixed variance and the equation's fundamental solution. De Bruijn's identity makes it general. As the heat equation spreads any density, that density's entropy increases at a rate equal to half its Fisher information. The prerequisites are only calculus and basic probability.

Contents

1 Introduction

Entropy is a number that measures uncertainty. Claude Shannon introduced it in 1948, in a paper on communication [Sha48], to make precise how much information a random source produces and how far its output can be compressed. The measure is built from the probabilities of the source alone, and it is not one choice among many. A few requirements that any measure of uncertainty ought to meet single out the formula, leaving free only the unit it is counted in.

This paper follows entropy from that setting to a place it is not obviously headed. Shannon's measure is made for a source with finitely many outcomes, a sum over a probability list. Replacing the sum with an integral carries it to a density on the real line, where it changes character and gives up properties the discrete version had. Restoring them calls for a change of object, from the entropy of one density to the relative entropy comparing two, and it is this comparison, not entropy itself, that survives the passage intact. Its local form is the Fisher information, and through that quantity entropy meets a classical differential equation. A density that spreads as time passes, ink diffusing through water, is governed by the heat equation, and as it spreads, its entropy climbs at a rate equal to half its Fisher information. A measure from communication theory and the equation for the flow of heat are joined in a single identity, due to de Bruijn. That identity is where the paper is headed.

The development is self-contained and asks only for calculus and basic probability. Section 2 states Shannon's problem and arrives at the entropy formula through a heuristic. Section 3 develops its basic properties and then shows the formula is forced, following a theorem of Khinchin that names four requirements and rules out every other function. Section 4 takes up two distributions at once, defining the relative entropy between them and the mutual information between two random variables. Section 5 crosses to the continuous line, where differential entropy takes on two defects, negativity and dependence on the coordinate. Section 6 shows that relative entropy crosses with neither, and turns it toward the relation with the heat equation.

2 Shannon's Problem

A random variable X is an outcome not yet observed, a draw that will land on one of several values, each with a fixed probability. We write capital X for the random draw itself and lowercase x for a particular value it can take. Here X takes values in a finite alphabet $\mathcal{X} = \{x_1, \ldots, x_n\}$, landing on x_i with probability p_i. Before we observe the outcome, we are uncertain. We want a single number that measures this uncertainty.

We ask three things of such a measure. Continuity: small changes in the probabilities produce small changes in the measure. Monotonicity at the uniform distribution: among n equally likely outcomes, uncertainty increases with n. Additivity: if X and Y are independent, the uncertainty of (X, Y) is the sum of the uncertainties of X and Y. These requirements point to the formula. The argument that follows is a heuristic; the precise statement, with the exact axioms it needs, is the uniqueness theorem of Section 3.

Assign a surprise $s(p)$ to an outcome with probability p. Take two independent outcomes, with probabilities p and q. Together they occur with probability pq. For the uncertainty of the pair to add, the surprise of both outcomes should be the sum of their separate surprises:

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

The logarithm does this. Setting $p = e^{-u}$ recasts the requirement $s(pq) = s(p) + s(q)$ as additivity in u, namely $s(e^{-(u+v)}) = s(e^{-u}) + s(e^{-v})$, and a continuous function additive in this way is linear in u, which forces $s(p) = c \ln p$ for some constant c. Monotonicity fixes the sign of c. Among n

equally likely outcomes, each has surprise $s(1/n) = -c \ln n$, and since every outcome carries the same surprise, the uncertainty is $-c \ln n$. For this to grow with n, we need $c < 0$. Setting $c = -1$ fixes the scale, and the surprise of an outcome of probability p is $-\ln p$. The surprise is non-negative on $(0, 1]$, but only because probabilities are at most 1. Densities can exceed 1, and in the continuous case both the surprise and the entropy built from it can be negative (Section 5).

The uncertainty of the whole distribution is the average surprise, each outcome weighted by how likely it is. A probability-weighted average is the expected value. For a function f of a random variable X,

$$\mathbb{E}[f(X)] = \sum_i p_i \, f(x_i).$$

The ordinary average $(a_1 + \cdots + a_n)/n$ is the special case where all outcomes are equally likely, each weight $1/n$. Applying this to surprise, the expected surprise of the distribution is:

Definition 2.1. The *entropy* of a discrete random variable X with probabilities $p_1, \ldots, p_n$ is

$$H(X) = \mathbb{E}[-\ln p(X)] = -\sum_{i=1}^{n} p_i \ln p_i.$$

Here $p(X)$ is the random variable equal to p_i when $X = x_i$. The surprise $-\ln p(X)$ is then itself random, and its expected value is $H(X)$.

We use the convention $0 \ln 0 = 0$, the limit of $x \ln x$ as $x \to 0^+$, so that outcomes of probability zero contribute nothing.

Throughout, the working base is the natural logarithm. Proofs and computations use ln and entropy comes out in nats, because the later sections differentiate and integrate these expressions and ln keeps the calculus clean. Bits (base 2) appear in a few examples, where we are counting symbols and a binary digit is the natural unit; the two differ only by a constant, and dividing a nat value by $\ln 2$ gives bits.

As a check, if X is deterministic (one outcome, probability 1), then $H(X) = -1 \cdot \ln 1 = 0$. There is no uncertainty and nothing to learn from observing the outcome.

Remark 2.2. Writing entropy as an average of per-outcome surprises already assumes the form $H = \sum_i p_i \, s(p_i)$, a constraint, not a convenience. The three properties above do not single out H on their own; what forces uniqueness is the grouping axiom (the conditional form of additivity), stated and used in Section 3.

The string `TTTTTTTT` from a source that almost always outputs `T` carries almost no information. Every symbol is expected, and the output compresses to almost nothing. The string `XTTTXTXT` carries more. Each `X` is surprising, and the sequence resists compression. Shannon's source coding theorem makes this precise.

Theorem 2.3 (Source coding theorem, informal)**.** *The entropy of X, in bits, is the smallest average number of binary digits per symbol any lossless code can achieve. Codes can approach this limit but never beat it.*

We state this without proof. The coding theory is its own subject.

The quantity $-\ln p(x)$ is the self-information of outcome x. It has several readings: the amount of information acquired by observing the outcome, how surprising it is, how many nats it costs to represent it. Entropy is the average self-information per symbol, and the source coding theorem says this average is exactly the compression limit.

Self-information and entropy depend on the probability distribution alone, not on what the outcomes mean. A random string of bits has maximum entropy and no meaning at all. Shannon

said as much in 1948: the "semantic aspects of communication are irrelevant to the engineering problem" [Sha48].

3 Properties and Uniqueness

Section 2 gave us the formula. Two questions remain. First, working only from the definition, what can we say about H? How large can it be, how small, and how does it behave when two random variables are combined? Second, is H the only function that could measure uncertainty, or one choice among many? Under the right axioms, the formula is forced, not chosen. The development follows Cover and Thomas [CT06].

3.1 Two bounds on entropy

Entropy is non-negative, and this is immediate from the definition.

Proposition 3.1. *$H(X) \geq 0$, with equality if and only if X is deterministic.*

Proof. Each probability p_i lies in $[0, 1]$, so $\ln p_i \leq 0$ and therefore $-p_i \ln p_i \geq 0$. The entropy is a sum of these terms, so $H(X) \geq 0$. For equality, every term must vanish. The term $-p_i \ln p_i$ is zero only when $p_i = 0$ or $p_i = 1$ (using $0 \ln 0 = 0$). Since the probabilities sum to 1, exactly one of them is 1 and the rest are 0. X takes a single value with certainty. □

The upper bound requires one inequality about the natural logarithm.

Lemma 3.2. *For every $x > 0$, $\ln x \leq x - 1$, with equality if and only if $x = 1$.*

Proof. Let $f(x) = x - 1 - \ln x$ on $(0, \infty)$. Then $f'(x) = 1 - 1/x$, which is negative for $x < 1$, zero at $x = 1$, and positive for $x > 1$. So f decreases then increases, with a global minimum at $x = 1$, where $f(1) = 0$. Hence $f(x) \geq 0$ for all $x > 0$, with equality only at $x = 1$. □

Among n outcomes, entropy is maximized when the distribution is uniform, yielding an upper bound of $\ln n$.

Proposition 3.3. *If X takes values in an alphabet of size n, then $H(X) \leq \ln n$, with equality if and only if X is uniform.*

Proof. Suppose first that $p_i > 0$ for every i. Apply Lemma 3.2 with $x = 1/(np_i)$:

$$\ln \frac{1}{np_i} \leq \frac{1}{np_i} - 1.$$

Multiply by p_i and sum over i:

$$\sum_i p_i \ln \frac{1}{np_i} \leq \sum_i \left(\frac{1}{n} - p_i \right) = 1 - 1 = 0.$$

Using $\ln \frac{1}{np_i} = -\ln n - \ln p_i$, the left-hand side is

$$\sum_i p_i \big(-\ln n - \ln p_i \big) = -\ln n \sum_i p_i - \sum_i p_i \ln p_i = -\ln n + H(X).$$

So $-\ln n + H(X) \leq 0$, that is $H(X) \leq \ln n$, with equality exactly when Lemma 3.2 holds with equality at every i, namely $1/(np_i) = 1$, i.e. $p_i = 1/n$ for all i, the uniform distribution. If instead

some $p_i = 0$, those outcomes contribute nothing to $H(X)$, so $H(X)$ is the entropy of X on its support, of size $m < n$. By the case just proved, $H(X) \leq \ln m < \ln n$ strictly, so equality forces every outcome to have positive probability. The bound holds in all cases, with equality only for the uniform distribution. □

So entropy lives in $[0, \ln n]$. The two endpoints are the extremes of uncertainty, 0 when one outcome is certain and $\ln n$ when all n are equally likely, and we know exactly which distribution sits at each end.

The probability simplex is the set of all distributions over n outcomes, vectors $(p_1, \ldots, p_n)$ with $p_i \geq 0$ and $\sum p_i = 1$. For $n = 2$ this is a line segment, for $n = 3$ a triangle. H is concave on this set, and this follows from the shape of each term in the sum.

Proposition 3.4. *H is a strictly concave function of $(p_1, \ldots, p_n)$ on the probability simplex.*

Proof. Write $H(p_1, \ldots, p_n) = \sum_{i=1}^{n} g(p_i)$ where $g(t) = -t \ln t$. Then $g'(t) = -\ln t - 1$, so $g''(t) = -1/t < 0$ for $t > 0$, and each g is strictly concave as a function of one variable. Take two distinct distributions $p \neq q$ and a weight $\lambda \in (0, 1)$. In each coordinate,

$$g\big(\lambda p_i + (1-\lambda) q_i\big) \geq \lambda\, g(p_i) + (1-\lambda)\, g(q_i),$$

with strict inequality whenever $p_i \neq q_i$. Since $p \neq q$, they differ in at least one coordinate, so summing over i makes the inequality strict:

$$H\big(\lambda p + (1-\lambda) q\big) > \lambda\, H(p) + (1-\lambda)\, H(q).$$

□

3.2 Joint and conditional entropy

So far we have worked with a single random variable. Two random variables bring joint and conditional distributions, and the question of how their combined uncertainty splits.

Consider two random variables X and Y with joint distribution $p(x, y)$.

Definition 3.5. The *joint entropy* of X and Y is

$$H(X, Y) = -\sum_{x} \sum_{y} p(x, y)\, \ln p(x, y) = \mathbb{E}[-\ln p(X, Y)].$$

This is the entropy of Section 2 applied to the pair (X, Y) treated as a single random variable on the product alphabet; the only change is that the formula now acts on the joint distribution $p(x, y)$ rather than a marginal.

The new idea is conditioning: fix a value x of X. Given $X = x$, the variable Y follows the conditional distribution $p(y \mid x)$, and its uncertainty is

$$H(Y \mid X = x) = -\sum_{y} p(y \mid x)\, \ln p(y \mid x).$$

This is one number for each value of x, the uncertainty about Y inside the slice where $X = x$. Different slices can look very different. In one, Y might be nearly deterministic; in another, close to uniform.

But X is itself random. Before observing it, we do not know which slice we will land in. The conditional entropy averages the per-slice entropies, weighting each by how likely its slice is.

Definition 3.6. The *conditional entropy* of Y given X is

$$H(Y \mid X) = \sum_x p(x)\, H(Y \mid X = x) = -\sum_x \sum_y p(x,y)\, \ln p(y \mid x) = \mathbb{E}[-\ln p(Y \mid X)].$$

These three quantities, $H(X)$, $H(Y \mid X)$, and $H(X,Y)$, satisfy an identity that reflects how joint probabilities decompose.

Proposition 3.7 (Chain rule)**.** $H(X,Y) = H(X) + H(Y \mid X)$.

Proof. The joint distribution factors as $p(x,y) = p(x)\, p(y \mid x)$. The logarithm converts this product into a sum:

$$\ln p(x,y) = \ln p(x) + \ln p(y \mid x).$$

Multiply both sides by $-p(x,y)$ and sum over all x and y:

$$-\sum_{x,y} p(x,y) \ln p(x,y) = -\sum_{x,y} p(x,y) \ln p(x) - \sum_{x,y} p(x,y) \ln p(y \mid x).$$

The left side is $H(X,Y)$. In the first sum on the right, $\ln p(x)$ does not depend on y, so summing $p(x,y)$ over y leaves the marginal $p(x)$, giving $-\sum_x p(x) \ln p(x) = H(X)$. The second sum is $H(Y \mid X)$ by definition. □

The total uncertainty in the pair is the uncertainty in X, plus whatever uncertainty about Y remains after X is known. This is what $H(Y \mid X)$ measures. If we are going to learn X but have not yet, how much uncertainty about Y will remain on average?

Reversing the roles gives $H(X,Y) = H(Y) + H(X \mid Y)$. The two decompositions agree because the joint entropy does not depend on which variable we account for first.

Conditioning can only reduce uncertainty, on average. The proof below uses only Lemma 3.2. Section 4 gives a shorter proof using relative entropy.

Proposition 3.8 (Conditioning reduces entropy)**.** $H(Y \mid X) \le H(Y)$*, with equality if and only if* X *and* Y *are independent.*

Proof. Expanding the definitions gives

$$H(Y) - H(Y \mid X) = -\sum_y p(y) \ln p(y) + \sum_{x,y} p(x,y) \ln p(y \mid x).$$

Since $p(y) = \sum_x p(x,y)$, the first sum can be written as a double sum $-\sum_{x,y} p(x,y) \ln p(y)$, so the two combine:

$$H(Y) - H(Y \mid X) = \sum_{x,y} p(x,y) \ln \frac{p(y \mid x)}{p(y)},$$

where the sum runs over pairs with $p(x,y) > 0$ (so that $p(y \mid x)$ and $p(y)$ are both positive). Apply Lemma 3.2 in the form $-\ln t \ge 1 - t$ with $t = p(y)/p(y \mid x)$:

$$\ln \frac{p(y \mid x)}{p(y)} = -\ln \frac{p(y)}{p(y \mid x)} \ge 1 - \frac{p(y)}{p(y \mid x)}.$$

Multiply by $p(x,y) \ge 0$ and sum:

$$H(Y) - H(Y \mid X) \ge \sum_{x,y} p(x,y) - \sum_{x,y} p(x,y) \cdot \frac{p(y)}{p(y \mid x)}.$$

The first sum equals 1. It runs over all pairs of positive joint probability, which carry the full mass. In the second, $p(x,y)/p(y \mid x) = p(x)$, so it equals $\sum_{(x,y):\, p(x,y)>0} p(x)\, p(y)$. Restoring the omitted pairs only adds non-negative terms, so this is at most $\sum_{x,y} p(x)\, p(y) = 1$. Hence

$$H(Y) - H(Y \mid X) \geq 1 - \sum_{(x,y):\, p(x,y)>0} p(x)\, p(y) \geq 1 - 1 = 0.$$

Equality forces two things at once: Lemma 3.2 must hold with equality on every pair with $p(x,y) > 0$, giving $p(y \mid x) = p(y)$ there; and the second sum must equal 1, which forces $p(x)\, p(y) = 0$ on every omitted pair. Together these say $p(x,y) = p(x)\, p(y)$ for all x, y, which is independence. For the converse, independence gives $p(y \mid x) = p(y)$ wherever $p(x,y) > 0$, so every log-ratio vanishes and $H(Y) - H(Y \mid X) = 0$. □

Learning the value of another variable cannot, on average, make you more uncertain. At worst it tells you nothing, and that happens exactly when the variables are independent.

3.3 Uniqueness

Everything in the last two subsections followed from the definition, treating the formula as given. But the formula was motivated by a heuristic in Section 2, and we noted there that the heuristic does not determine it uniquely. The natural question is whether there exist axioms that do, conditions natural enough to accept as requirements on any measure of uncertainty, and strong enough to force the formula. Khinchin [Khi57] proved that four such axioms suffice. We state them for a function F defined on every finite probability distribution $(p_1, \ldots, p_n)$ with $n \geq 1$.

1. **Continuity.** $F(p_1, \ldots, p_n)$ is continuous in the probabilities.

2. **Maximum at uniform.** For each n, F is maximized by the uniform distribution:

$$F(p_1, \ldots, p_n) \leq F(1/n, \ldots, 1/n).$$

3. **Expansibility.** Adding an impossible outcome changes nothing:

$$F(p_1, \ldots, p_n, 0) = F(p_1, \ldots, p_n).$$

4. **Grouping.** If two outcomes with probabilities p_1 and p_2 are merged into a single outcome of probability $p_1 + p_2$, then

$$F(p_1, p_2, p_3, \ldots, p_n) = F(p_1 + p_2,\, p_3, \ldots, p_n) \;+\; (p_1 + p_2)\, F\left(\frac{p_1}{p_1 + p_2},\; \frac{p_2}{p_1 + p_2}\right).$$

A convention underlies all four. F is a function of the distribution itself, unchanged if we reorder the outcomes. This is what it means to measure the uncertainty of a distribution. The labels carry no information. This convention is built into Khinchin's original formulation [Khi57]. It lets us state the grouping axiom for the first two outcomes without loss of generality, since any two outcomes can be merged after relabeling. (The grouping axiom in this two-outcome form is due to Faddeev, whose formulation lists this label-independence as an explicit symmetry axiom and manages without maximum-at-uniform, at the price of a much harder proof. Khinchin's own fourth axiom asserts the decomposition directly for a pair of variables: the entropy of the joint distribution is the entropy of

the first variable plus the average entropy of the second given the first. That is the partition form derived below, so the two systems meet there, and the proof that follows is Khinchin's.)

The first three axioms need little justification. A reasonable measure of uncertainty should vary smoothly, peak when all outcomes are equally likely, and ignore outcomes that cannot occur. These are not the three properties of Section 2, but they line up with them. Continuity is the same. Grouping is the conditional form of additivity (Remark 2.2). Max-at-uniform and expansibility together recover monotonicity, as Stage 1 shows. Expansibility is the one new requirement.

The fourth axiom is the structural one. It describes what happens when two outcomes are lumped together. Suppose outcomes 1 and 2 are merged into a single composite outcome with probability $p_1 + p_2$, reducing the distribution from n outcomes to $n-1$. The axiom says the entropy of the original distribution is the sum of two terms. The first term, $F(p_1 + p_2,\, p_3, \ldots, p_n)$, is the entropy of the coarser distribution where outcomes 1 and 2 have been collapsed into one. The second term is a correction for the information the collapse destroyed. Given that the composite outcome occurred (probability $p_1 + p_2$), how uncertain are we about which of the two original outcomes it was? That within-group uncertainty is $F(p_1/(p_1+p_2),\ p_2/(p_1+p_2))$, weighted by the probability $p_1 + p_2$ that the group occurs at all. The total entropy decomposes exactly into the coarse-level uncertainty plus the expected fine-level uncertainty.

As a concrete instance, consider three outcomes, rain (0.3), snow (0.2), and sun (0.5). Merge rain and snow into a single outcome "precipitation" with probability 0.5. The axiom says

$$F(0.3,\, 0.2,\, 0.5) \;=\; F(0.5,\, 0.5) \;+\; 0.5 \cdot F(0.6,\, 0.4).$$

The first term is the entropy of the coarse question, precipitation or sun. The second is the entropy of the follow-up, rain ($0.3/0.5 = 0.6$) or snow ($0.2/0.5 = 0.4$) given that it precipitates. The factor 0.5 is there because the follow-up question only arises when it precipitates. The division by $p_1 + p_2$ inside F is conditioning, restricting the distribution to the group.

Applying axiom (4) repeatedly extends it to any number of outcomes being merged. To see how, take $k = 3$ with $q = p_1 + p_2 + p_3$. Apply the axiom twice, first merging outcomes 1 and 2, then merging the result with outcome 3:

$$\begin{aligned} F(p_1, p_2, p_3, \ldots, p_n) &= F(p_1{+}p_2,\, p_3, \ldots, p_n) \\ &\qquad + (p_1{+}p_2)\, F\Big(\tfrac{p_1}{p_1+p_2},\, \tfrac{p_2}{p_1+p_2}\Big) \\ &= F(q,\, p_4, \ldots, p_n) + q\, F\Big(\tfrac{p_1+p_2}{q},\, \tfrac{p_3}{q}\Big) \\ &\qquad + (p_1{+}p_2)\, F\Big(\tfrac{p_1}{p_1+p_2},\, \tfrac{p_2}{p_1+p_2}\Big). \end{aligned}$$

The last two terms combine into $q\, F(p_1/q,\, p_2/q,\, p_3/q)$. To verify, apply axiom (4) to this last F, merging its first two arguments:

$$\begin{aligned} F\Big(\tfrac{p_1}{q},\, \tfrac{p_2}{q},\, \tfrac{p_3}{q}\Big) &= F\Big(\tfrac{p_1+p_2}{q},\, \tfrac{p_3}{q}\Big) \\ &\qquad + \tfrac{p_1+p_2}{q}\, F\Big(\tfrac{p_1}{p_1+p_2},\, \tfrac{p_2}{p_1+p_2}\Big). \end{aligned}$$

Multiply through by q and compare. The general identity follows by induction on the group size k. The base case $k = 2$ is axiom (4) itself. Suppose the identity holds for groups of size k, and take a group of size $k+1$ with total $q = p_1 + \cdots + p_{k+1}$. Apply the size-$k$ case to the first k outcomes, then merge the resulting composite with outcome $k+1$ by axiom (4), and recombine the two within-group

terms as above. This gives the identity for $k+1$. The computation just shown is exactly this step at $k=2$. So if outcomes 1 through k form a group with total probability $q = p_1 + \cdots + p_k$, then

$$F(p_1, \ldots, p_n) = F(q,\, p_{k+1}, \ldots, p_n) + q\, F(p_1/q, \ldots, p_k/q). \tag{1}$$

For a partition into several groups, apply (1) to each group in turn. The entropy of the whole distribution equals the entropy of the group-level distribution plus the weighted average of the within-group entropies.

This is the axiom that forces uniqueness. The three properties from Section 2 (continuity, monotonicity at uniform, additivity for independent variables) do not suffice. The Rényi entropies [Ren61] satisfy all three for every order $\alpha > 0$, and only $\alpha = 1$ is Shannon's. The grouping axiom excludes every $\alpha \neq 1$. The Rényi entropies are additive for independent variables but do not decompose recursively under grouping.

Theorem 3.9 (Khinchin)**.** *If F satisfies axioms* (1)–(4)*, then*

$$F(p_1, \ldots, p_n) = -K \sum_{i=1}^{n} p_i \ln p_i$$

for some constant $K \geq 0$.

If F is not identically zero, then $K > 0$, and the constant is the choice of unit. Setting $K = 1$ recovers the entropy of Section 2.

Proof. Define $L(n) = F(1/n, \ldots, 1/n)$, the value of F on the uniform distribution over n outcomes.

Stage 1. We show $L(mn) = L(m) + L(n)$.

Take mn equally likely outcomes, each with probability $1/(mn)$, and partition them into m groups of n. In the notation of (1), each group has $k = n$ outcomes, total probability $q = n \cdot \frac{1}{mn} = \frac{1}{m}$, and within-group probabilities $p_i/q = \frac{1/(mn)}{1/m} = \frac{1}{n}$. Applying (1) to all m groups:

$$L(mn) = F\left(\frac{1}{m}, \ldots, \frac{1}{m}\right) + \sum_{j=1}^{m} \frac{1}{m}\, F\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = L(m) + m \cdot \frac{1}{m} \cdot L(n) = L(m) + L(n).$$

Note also that L is non-decreasing. By expansibility, $L(n) = F(1/n, \ldots, 1/n, 0)$, and by the maximum-at-uniform axiom applied to the $(n+1)$-outcome distribution $(1/n, \ldots, 1/n, 0)$, this is at most $F(1/(n{+}1), \ldots, 1/(n{+}1)) = L(n+1)$.

Finally, setting $n = 1$ in $L(m \cdot 1) = L(m) + L(1)$ gives $L(1) = 0$.

Stage 2. We show $L(n) = K \ln n$ for some $K \geq 0$.

The equation $L(mn) = L(m) + L(n)$ says L converts products to sums, and among non-decreasing functions on the positive integers the logarithm is the only one that does. The argument below makes this precise by showing $L(n)/\ln n$ is the same for every $n \geq 2$. It squeezes powers r^m and s^n together as integers, computes L of both by multiplicativity, and lets monotonicity force the ratio $L(s)/L(r)$.

The argument below divides by $L(r)$, so we need $L(r) > 0$ for $r \geq 2$. If L is identically zero, set $K = 0$ and we are done. If not, then since L is non-negative (L is non-decreasing and $L(1) = 0$), $L(r) > 0$ for at least one $r \geq 2$. We claim more strongly that $L(r) > 0$ for *every* $r \geq 2$. Suppose for the sake of contradiction that $L(r_0) = 0$ for some $r_0 \geq 2$. Multiplicativity gives $L(r_0^k) = k\, L(r_0) = 0$ for all k. Every integer s sits below some power r_0^k, and L is non-decreasing with $L(1) = 0$, so

$L(s) = 0$ for all s, but L is not identically zero, a contradiction. The supposition $L(r_0) = 0$ is therefore false, and $L(r) > 0$ for every $r \geq 2$.

Fix integers $r \geq 2$ and $s \geq 2$. For any positive integer n, choose m by the condition $r^m \leq s^n < r^{m+1}$. Taking logarithms gives

$$\frac{m}{n} \leq \frac{\ln s}{\ln r} < \frac{m}{n} + \frac{1}{n}. \tag{2}$$

Since L is non-decreasing, $L(r^m) \leq L(s^n) \leq L(r^{m+1})$. Multiplicativity gives $L(r^m) = m\,L(r)$ and $L(s^n) = n\,L(s)$, so

$$\frac{m}{n} \leq \frac{L(s)}{L(r)} \leq \frac{m}{n} + \frac{1}{n}. \tag{3}$$

Combining (2) and (3):

$$\left| \frac{L(s)}{L(r)} - \frac{\ln s}{\ln r} \right| \leq \frac{1}{n}.$$

The left side is a fixed number, the bound $1/n$ is arbitrarily small, and the only non-negative number smaller than every $1/n$ is zero. The two ratios are equal. Since r and s are arbitrary, $L(n)/\ln n$ takes the same value for all $n \geq 2$. Call it K. Then $L(n) = K \ln n$ for $n \geq 2$, and the formula holds at $n = 1$ as well since both sides are zero. The constant K is non-negative because L is non-decreasing.

Stage 3. We extend from uniform to arbitrary distributions.

By continuity, it suffices to handle rational probabilities $p_i = k_i/N$ where each k_i is a positive integer and $\sum_i k_i = N$. Take N equally likely outcomes and group them into n clusters of sizes $k_1, \ldots, k_n$. The j-th cluster has total probability $k_j/N = p_j$, and within it the conditional distribution is uniform on k_j outcomes. By the grouping axiom,

$$L(N) = F(p_1, \ldots, p_n) + \sum_{j=1}^{n} p_j\, L(k_j).$$

Since $L(N) = K \ln N$ and $L(k_j) = K \ln k_j$,

$$K \ln N = F(p_1, \ldots, p_n) + K \sum_j p_j \ln k_j.$$

Writing $k_j = p_j N$ and solving for F:

$$\begin{aligned} F(p_1, \ldots, p_n) &= K \ln N - K \sum_j p_j \ln(p_j N) \\ &= K \ln N - K \sum_j p_j (\ln p_j + \ln N) \\ &= K \ln N - K \ln N \sum_j p_j - K \sum_j p_j \ln p_j \\ &= -K \sum_j p_j \ln p_j, \end{aligned}$$

using $\sum p_j = 1$. This proves the formula for rational distributions. The rationals are dense. Every distribution $(p_1, \ldots, p_n)$ is a limit of distributions with rational entries. Both sides of $F = -K \sum_i p_i \ln p_i$ are continuous in the p_i, the left by axiom (1) and the right because $t \ln t$ is continuous, so the identity passes to the limit and holds for every distribution. If some $p_i = 0$, apply expansibility to drop that outcome and use the formula on the remaining positive probabilities. $\square$

The four axioms leave no freedom except the unit of measurement. The heuristic in Section 2 assumed entropy has the form $\sum p_i\, s(p_i)$, a sum of per-outcome contributions, and derived $s(p) = -\ln p$. But why should a measure of uncertainty decompose per-outcome at all? The grouping axiom answers this. It speaks only of how entropy splits across groups, yet Stage 3 shows it forces F into exactly the form $-K \sum p_i \ln p_i$. The per-outcome structure is therefore derived, not assumed.

3.4 Examples

The simplest non-trivial distribution is Bernoulli, two outcomes with probabilities p and $1 - p$. Its entropy, written $H(p)$ and called the *binary entropy function*, is

$$H(p) = -p \ln p - (1 - p) \ln(1 - p)$$

and is zero at the endpoints $p = 0$ and $p = 1$ (deterministic) and reaches its maximum $H(1/2) = \ln 2 \approx 0.693$ nats at $p = 1/2$ (uniform on two outcomes). The function is symmetric about $1/2$ and strictly concave.

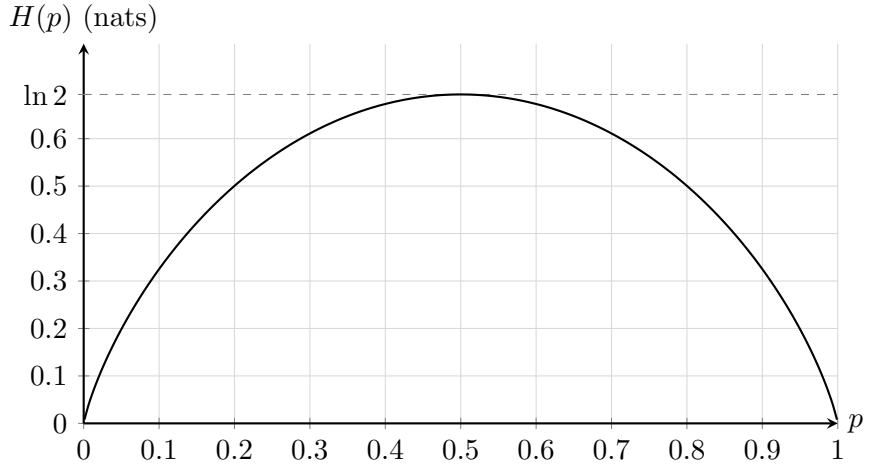

Figure 1: The binary entropy function $H(p)$, the entropy of a Bernoulli(p) variable.

For a fair six-sided die, all faces are equally likely and $H = \ln 6 \approx 1.79$ nats, the maximum for six outcomes. A loaded die with probabilities (1/2, 1/6, 1/12, 1/12, 1/12, 1/12) has

$$H = \tfrac{1}{2} \ln 2 + \tfrac{1}{6} \ln 6 + \tfrac{1}{3} \ln 12 \approx 1.47 \text{ nats},$$

strictly below $\ln 6$. The gap measures the departure from uniformity, and Proposition 3.3 guarantees that it is non-negative, vanishing only for the fair die.

A larger example is the 26 letters of English. If letters were equally frequent, the entropy would be $\log_2 26 \approx 4.70$ bits per letter (dividing the nat value by $\ln 2$ converts units). From the actual single-letter frequencies of English text, where `E` appears about 12.7% of the time and `Z` about 0.07%, Shannon [Sha51] computed the entropy at 4.14 bits per letter. The shortfall of roughly 0.6 bits reflects the non-uniformity of letter usage. A few letters carry most of the weight, and many contribute almost nothing.

But single-letter frequencies capture only part of the picture. English letters are not independent. After `Q` the next letter is almost certainly `U`, after `T` it is often `H`. Knowing the previous letter helps predict the next, so the conditional entropy $H(X_2 \mid X_1)$ is less than $H(X_1)$. More context helps

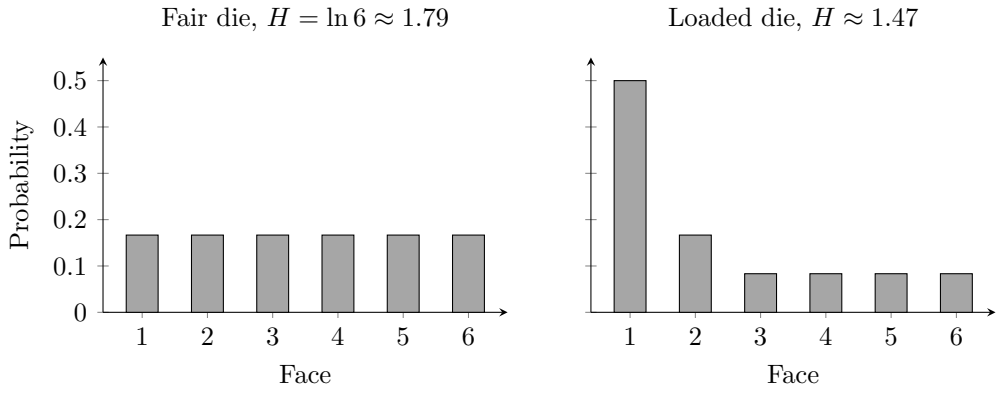

Figure 2: A fair die and a loaded die over the six faces. Even probabilities are hardest to predict and give the most entropy; concentrating mass on face 1 lowers it.

further. $H(X_3 \mid X_1, X_2)$ drops again, and each additional letter of history makes the next letter more predictable. The limit of $H(X_n \mid X_1, \ldots, X_{n-1})$ as $n \to \infty$ is the *entropy rate*, the per-letter uncertainty once all the structure of the language (letter patterns, word boundaries, grammar) is accounted for.

Shannon estimated this entropy rate in his 1951 paper [Sha51]. He showed human subjects text one character at a time and asked them to guess the next character. Characters guessed correctly on the first try carried little uncertainty; those requiring many guesses carried a lot. From the guess statistics, he bounded the entropy rate at roughly 1 bit per letter, far below the single-letter value of 4.14.

One bit is the answer to a single yes-or-no question. Shannon's result says that, on average, one binary question is enough to determine the next letter of English text, from an alphabet of 26. A uniform alphabet would need nearly 5 bits per letter. English needs about 1, compressing away close to 4 bits of redundancy per letter. Apparently English is tractable to learn.

4 Relative Entropy and Mutual Information

Sections 2 and 3 measured the uncertainty of a single distribution. This section takes up two. Given p and q on the same alphabet, how far apart are they?

4.1 Relative entropy

To compare two distributions, work outcome by outcome. At outcome i the two distributions assign probabilities p_i and q_i, and the comparison is their ratio p_i/q_i. The ratio is 1 where they agree, greater than 1 where q falls short of p, and less than 1 where q exceeds it. A difference $p_i - q_i$ would conflate errors of very different significance. If p assigns 0.50 to an outcome and q assigns 0.51, the difference is 0.01, and q is off by two percent. If p assigns 0.001 and q assigns 0.011, the difference is again 0.01, but now q makes the outcome eleven times likelier than it is. The ratio separates these, about 0.98 in the first case and about 0.09 in the second.

This is one number per outcome. To reduce them to a single measure of the discrepancy between q and p, take the logarithm, as in Section 2, and average against p. The logarithm makes the measure additive over independent sources. The average is weighted by p because p governs which outcomes occur. The result is the mean log-ratio $\sum_i p_i \ln(p_i/q_i)$.

Definition 4.1. The *relative entropy*, or *Kullback–Leibler divergence* [KL51], of p from q is

$$D(p \,\|\, q) = \sum_i p_i \ln \frac{p_i}{q_i},$$

summed over the outcomes with $p_i > 0$. We use the conventions $0 \ln \frac{0}{q} = 0$ and $p_i \ln \frac{p_i}{0} = +\infty$ for $p_i > 0$.

The first convention says that outcomes the true distribution never produces contribute nothing, since they never occur. The second says that if p gives positive probability to an outcome that q rules out, the divergence is infinite. q has assigned zero probability to something that occurs.

The log-ratio is also a difference of surprises. From Section 2, an outcome of probability r carries surprise $-\ln r$ to anyone who expects it, so

$$\ln \frac{p_i}{q_i} = (-\ln q_i) - (-\ln p_i)$$

is the gap between the surprise outcome i causes someone who believes q and the surprise it causes someone who believes the truth p. Averaged against p, the divergence $D(p \,\|\, q)$ is the mean surprise of believing q, namely $\sum_i p_i(-\ln q_i)$, minus the mean surprise $H(p)$ of believing the truth.

A measure of difference should have two properties: it should never be negative, since nothing is closer to p than p itself, and it should be zero exactly when $q = p$, since a distribution has no discrepancy from itself. A function of two distributions with these two properties is called a *divergence*, a weaker notion than a distance. It need not be symmetric, and need not satisfy the triangle inequality.

Neither property is visible in the formula. The terms $p_i \ln(p_i/q_i)$ are positive where q underestimates an outcome but negative where q overestimates it, and it is not obvious that the positive terms always win. That they do is the content of the following proposition.

Proposition 4.2 (Gibbs' inequality)**.** *$D(p \,\|\, q) \geq 0$, with equality if and only if $p = q$.*

Proof. If $D(p\|q) = +\infty$ there is nothing to prove, so assume it is finite, which means $q_i > 0$ wherever $p_i > 0$. Apply Lemma 3.2 in the form $\ln t \leq t - 1$ with $t = q_i/p_i$:

$$-D(p\|q) = \sum_{i:\, p_i>0} p_i \ln \frac{q_i}{p_i} \leq \sum_{i:\, p_i>0} p_i \left(\frac{q_i}{p_i} - 1 \right) = \sum_{i:\, p_i>0} q_i - 1 \leq 0,$$

where the last step uses $\sum_{i:\, p_i>0} q_i \leq \sum_i q_i = 1$. Hence $D(p\|q) \geq 0$. Equality needs both inequalities tight at once. The first forces $q_i/p_i = 1$ at every i with $p_i > 0$, so $q_i = p_i$ on the support of p. The second forces $\sum_{i:\, p_i>0} q_i = 1$, so q places no mass outside that support. Together they give $p_i = q_i$ for every i. □

The positive terms always win.

The upper bound $H(X) \leq \ln n$ from Section 3 is now a corollary. Take q to be the uniform distribution $u = (1/n, \ldots, 1/n)$. Then

$$D(p \,\|\, u) = \sum_i p_i \ln \frac{p_i}{1/n} = \sum_i p_i \ln p_i + \ln n \sum_i p_i = \ln n - H(X),$$

and $D(p\|u) \geq 0$ gives $H(X) \leq \ln n$, with equality when $p = u$.

Rearranged, the same identity reads

$$H(X) = \ln n - D(p \,\|\, u).$$

The entropy of p is its divergence from the uniform distribution, subtracted from the maximum $\ln n$. A distribution near uniform has high entropy, a concentrated one far from uniform has low entropy. Entropy is not an isolated quantity. It measures p against a reference, and on a finite alphabet that reference is the uniform distribution, so natural a choice that it goes unnoticed. The uniform reference contributes only the constant $\ln n$, the same for every p, since $-\sum_i p_i \ln(1/n) = \ln n$ whatever p is. It leaves no mark on the formula $H(X) = -\sum_i p_i \ln p_i$, and one can write entropy for a whole section, as we have, without seeing the comparison folded inside. Section 5 removes the disguise. On the continuous line there is no uniform distribution to hide behind, and the reference resurfaces, no longer a constant absorbed into the formula but a divergent term that must be subtracted off by hand.

KL is not symmetric. The quantity $D(p\|q)$ asks how costly it is to treat the truth p as though it were q; the reversed quantity $D(q\|p)$ asks the opposite, how costly it is to treat the truth q as though it were p. These are different questions, because underestimating an outcome and overestimating it are penalized differently. A small example shows the gap. Take $p = (\frac{1}{2}, \frac{1}{2})$ and $q = (\frac{1}{10}, \frac{9}{10})$ on two outcomes:

$$D(p \,\|\, q) = \tfrac{1}{2} \ln 5 + \tfrac{1}{2} \ln \tfrac{5}{9} = \ln \tfrac{5}{3} \approx 0.511 \text{ nats},$$

$$D(q \,\|\, p) = \tfrac{1}{10} \ln \tfrac{1}{5} + \tfrac{9}{10} \ln \tfrac{9}{5} = \tfrac{9}{5} \ln 3 - \ln 5 \approx 0.368 \text{ nats}.$$

Both are positive, as Gibbs requires, and they are unequal. The asymmetry grows without bound at the edge. If p places mass where q assigns none, $D(p\|q) = \infty$, while $D(q\|p)$ stays finite because q never asks to encode the offending outcome.

The divergence also fails the triangle inequality, and symmetrizing it does not change this, so it is not a metric, a rule for measuring distance in a space. In its local behavior it resembles a squared distance more than a distance. As q approaches p, the divergence shrinks like the square of the gap between them. Section 6.3 makes this precise for q a small shift of p, where the coefficient is the Fisher information, the quantity that ties entropy to the heat equation.

4.2 Mutual information

Relative entropy compares any two distributions on the same alphabet. A particular choice of the two measures something Section 3 could not, the dependence between two random variables X and Y.

Independence is a statement about distributions. X and Y are independent exactly when their joint distribution is the product of the marginals, $p(x, y) = p(x)\, p(y)$. When they are dependent, the joint and the product differ, and the divergence between them measures by how much. Comparing the true joint $p(x, y)$ against the product $p(x)\, p(y)$ gives the definition.

Definition 4.3. The *mutual information* of X and Y is

$$I(X;Y) = D\big(p(x,y) \,\|\, p(x)p(y)\big) = \sum_{x,y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)}.$$

The product $p(x)p(y)$ is itself a distribution on pairs, so this is a genuine relative entropy, the cost of modeling X and Y as independent when they are not. If they truly are independent, the joint equals the product, the divergence is zero, and $I(X;Y) = 0$.

The definition reads I as a divergence between two-dimensional distributions. It also equals combinations of the one-dimensional entropies from Section 3.

Proposition 4.4. $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = H(X) + H(Y) - H(X,Y)$.

Proof. Factor the joint as $p(x,y) = p(x)\,p(y \mid x)$, so that $\frac{p(x,y)}{p(x)p(y)} = \frac{p(y|x)}{p(y)}$. Then

$$I(X;Y) = \sum_{x,y} p(x,y) \ln \frac{p(y \mid x)}{p(y)} = \sum_{x,y} p(x,y) \ln p(y \mid x) - \sum_{x,y} p(x,y) \ln p(y).$$

The first sum is $-H(Y \mid X)$ by definition. In the second, $\ln p(y)$ does not depend on x, so summing $p(x,y)$ over x leaves $p(y)$, and the sum is $\sum_y p(y) \ln p(y) = -H(Y)$. Hence $I(X;Y) = H(Y) - H(Y \mid X)$. Factoring the other way, $p(x,y) = p(y)\,p(x \mid y)$, gives $I(X;Y) = H(X) - H(X \mid Y)$ by the same steps. For the third form, the chain rule (Proposition 3.7) gives $H(X \mid Y) = H(X,Y) - H(Y)$; substituting into $I(X;Y) = H(X) - H(X \mid Y)$ leaves $H(X) + H(Y) - H(X,Y)$. □

Each form reads differently. $H(X) - H(X \mid Y)$ is the amount that learning Y cuts the uncertainty in X. $H(Y) - H(Y \mid X)$ says the cut is the same the other way, that learning X reduces the uncertainty in Y by an equal amount, which is why the quantity is called mutual. $H(X) + H(Y) - H(X,Y)$ is symmetric on its face. The shared information is what is left after the joint uncertainty $H(X,Y)$ is taken out of the total $H(X) + H(Y)$. Figure 3 holds all three at once.

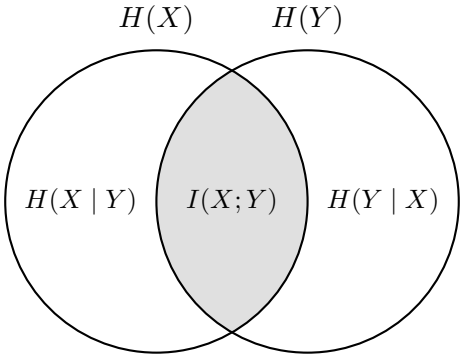

Figure 3: Areas stand for entropies. The two circles are $H(X)$ and $H(Y)$, their union is the joint entropy $H(X,Y)$, and the overlap is the mutual information $I(X;Y)$. Each crescent is a conditional entropy, what is left of one variable's uncertainty once the other is known.

Setting $Y = X$ gives a special case. Knowing X leaves no uncertainty about X, so $H(X \mid X) = 0$, and the first identity gives $I(X;X) = H(X)$. A variable carries exactly as much information about itself as it had uncertainty to begin with.

Because $I(X;Y)$ is a relative entropy, Gibbs applies directly.

Proposition 4.5. *$I(X;Y) \geq 0$, with equality if and only if X and Y are independent.*

Proof. Apply Proposition 4.2 to $p(x,y)$ and $p(x)p(y)$. Equality holds exactly when the two distributions coincide, that is $p(x,y) = p(x)p(y)$ for all x, y, which is independence. □

This recovers Proposition 3.8. Combining $I(X;Y) \geq 0$ with $I(X;Y) = H(Y) - H(Y \mid X)$ yields $H(Y \mid X) \leq H(Y)$, and the condition for equality, independence, now falls out of Gibbs.

Entropy measures the uncertainty in one distribution, relative entropy the gap between two, mutual information the dependence between two variables. The next two sections carry all three to densities on the real line, where the sums become integrals and not everything survives the passage.

5 Differential Entropy

A continuous random variable has no finite list of outcomes to sum over. It has a density $p(x)$, and extending entropy to this case means deciding what replaces the sum $-\sum_i p_i \ln p_i$. The natural guess is the integral $-\int p \ln p \, dx$. The guess is almost right. Discretizing the line and refining the bins does produce this integral, but alongside a term that grows without bound, and the integral is only the finite part that remains once that term is removed.

5.1 The discretization limit

A continuous variable can be approximated by a discrete one. Split the line into intervals of width Δ and record which interval X lands in, keeping the interval and discarding the position inside it. This record is an ordinary discrete variable, with the entropy of Section 2, and narrowing the intervals sharpens the approximation.

Fix a width $\Delta > 0$ and let $B_i = [i\Delta, (i+1)\Delta)$ be the intervals, or bins. The variable X_Δ records which bin contains X. It equals i with probability

$$p_i = \int_{B_i} p(x)\, dx.$$

Since X_Δ is discrete, its entropy is the one from Section 2,

$$H(X_\Delta) = -\sum_i p_i \ln p_i,$$

now a series over countably many bins rather than a finite sum. For a density with heavy enough tails it can diverge, so we assume throughout this section that p is continuous and decays fast enough for this series and the integrals below to converge; a bounded density with finite variance is enough. We let $\Delta \to 0$.

When Δ is small, p changes little across a single bin, so the probability in a bin is close to the density there times the width, $p_i \approx p(x_i)\Delta$ for a point x_i in the bin. This step can be made exact. The probability $p_i = \int_{B_i} p\, dx$ is the area under the density across the bin, and any area of this kind equals the average height of the curve times the width Δ. The mean value theorem for integrals says a continuous density attains that average height at some point ξ_i inside the bin,[1] so p_i is exactly that height times the width:

$$p_i = \int_{B_i} p(x)\, dx = p(\xi_i)\, \Delta.$$

The point ξ_i is only known to lie somewhere in B_i. Its exact location is never used. Substitute $p_i = p(\xi_i)\Delta$ into the entropy, both in the factor out front and inside the logarithm:

$$H(X_\Delta) = -\sum_i p_i \ln p_i = -\sum_i p(\xi_i)\Delta \, \ln\big(p(\xi_i)\Delta\big).$$

A logarithm of a product is a sum of logarithms, $\ln\big(p(\xi_i)\Delta\big) = \ln p(\xi_i) + \ln \Delta$, so

$$H(X_\Delta) = -\sum_i p(\xi_i)\Delta \, \big(\ln p(\xi_i) + \ln \Delta\big).$$

[1] A continuous function on a closed interval equals its average value there at some interior point. The average of p over B_i is $\frac{1}{\Delta}\int_{B_i} p\, dx$, so $\int_{B_i} p\, dx = p(\xi_i)\, \Delta$.

Expanding the bracket splits this into two sums,

$$H(X_\Delta) = -\sum_i p(\xi_i)\Delta\, \ln p(\xi_i) \;-\; \sum_i p(\xi_i)\Delta\, \ln \Delta,$$

and in the second sum $\ln \Delta$ is the same in every term, so it comes out of the sum:

$$H(X_\Delta) = -\sum_i p(\xi_i) \ln p(\xi_i)\, \Delta \;-\; \ln \Delta \sum_i p(\xi_i)\Delta.$$

Take the two sums one at a time.

The second sum is exact. Each $p(\xi_i)\Delta$ is again p_i, and the bins cover the line without overlap, so every value of X falls in exactly one of them and the p_i add to 1:

$$\sum_i p(\xi_i)\Delta = \sum_i p_i = 1.$$

The second term is therefore $-\ln \Delta = \ln(1/\Delta)$, with no approximation.

The first sum is a Riemann sum for $-\int p \ln p\, dx$. Each term is the value of $-p \ln p$ at the sample point ξ_i times the width Δ, so the sum is the combined area of rectangles fitted to the curve. As the bins narrow the rectangles track the curve more closely, and the sum converges to the integral as $\Delta \to 0$,

$$-\sum_i p(\xi_i) \ln p(\xi_i)\, \Delta \;\longrightarrow\; -\int p(x) \ln p(x)\, dx,$$

provided p is Riemann integrable and decays fast enough at the tails for the integral over the whole line to converge; the precise statement is Theorem 8.3.1 of Cover and Thomas [CT06].

Write $h(X) = -\int p \ln p\, dx$ for this limit. Combining the two terms,

$$H(X_\Delta) = h(X) + \ln \frac{1}{\Delta} + o(1) \qquad (\Delta \to 0),$$

where $o(1)$ denotes the gap between the Riemann sum and its limit. That gap is the only approximation anywhere in the derivation, and it vanishes as $\Delta \to 0$. Figure 4 shows a standard normal under this binning at two widths, with $H(X_\Delta)$ rising as the bins narrow.

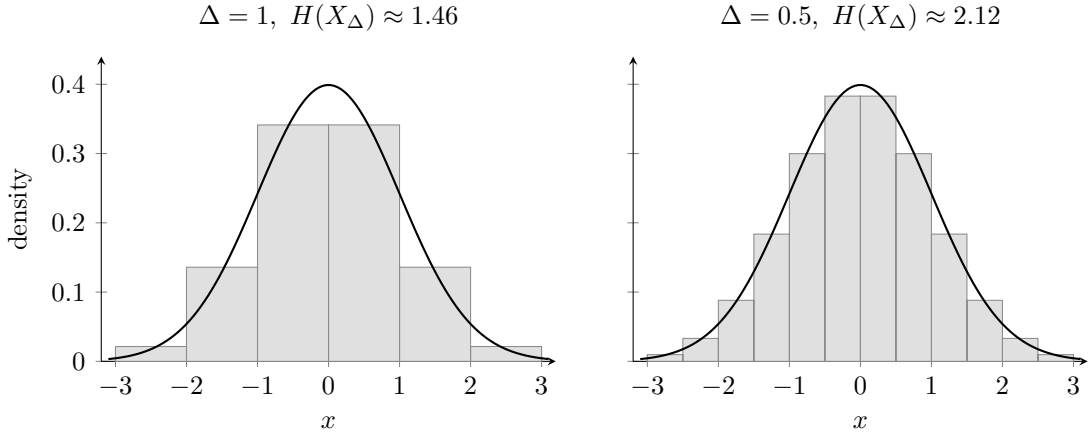

Figure 4: A standard normal (curve) and its approximation by bins of width $\Delta = 1$ on the left and $\Delta = 0.5$ on the right. Each bar has height p_i/Δ, so its area is the bin probability p_i and the bars are a piecewise-constant approximation to the density. The finer grid tracks the curve more closely, and its discrete entropy $H(X_\Delta)$ is larger.

The term $\ln(1/\Delta)$ grows without bound as $\Delta \to 0$, so $H(X_\Delta) \to \infty$. This divergence is not an error in the calculation: pinning down a real number exactly takes an infinite amount of information. In bits, $\log_2(1/\Delta)$ is the number of binary digits needed to identify which bin contains X, and each halving of Δ adds one more digit, hence one more bit. A real number has infinitely many digits, so the uncertainty in its exact value is infinite. The integral $-\int p \ln p$ is not that uncertainty. It is the finite part that remains after the divergent term $\ln(1/\Delta)$ is subtracted.

Definition 5.1. The *differential entropy* of a continuous random variable X with density p is

$$h(X) = \mathbb{E}[-\ln p(X)] = -\int p(x) \ln p(x)\, dx,$$

when the integral exists. We keep the convention $0 \ln 0 = 0$.

As in the discrete case, $h(X)$ is an average. For a continuous X with density p, the expectation of a function g is the density-weighted integral

$$\mathbb{E}[g(X)] = \int g(x)\, p(x)\, dx,$$

the continuous version of the weighted sum $\sum_i p_i\, g(x_i)$, each value x weighted by the density $p(x)$ there. Differential entropy is this average for $g = -\ln p$. Capital X is the random draw and lowercase x a value it can land on, so $-\ln p(X)$, the surprise of wherever the draw falls, is itself random, and $h(X)$ is its average over all draws.

The term $\ln(1/\Delta)$ is the same for every density. It depends only on Δ, while the density enters only through $h(X)$.

Three things follow from this, the first two in the next section. First, h need not be positive. It is what remains after an infinite quantity is subtracted, so it can be negative. Second, the resolution Δ is a length on the x-axis, so rescaling or reparametrizing x changes what is subtracted, and h changes with it, unlike the discrete entropy, which is unchanged by relabeling outcomes. Third (Section 6), two densities compared at the same resolution share the same $\ln(1/\Delta)$, and it cancels in the comparison. This is why relative entropy carries over to the continuous case unchanged while h does not.

5.2 Negativity and coordinate dependence

The first two are concrete properties of h, both visible on simple densities.

Take X uniform on $[0, a]$, with density $p(x) = 1/a$ on the interval and 0 outside. Then

$$h(X) = -\int_0^a \frac{1}{a} \ln \frac{1}{a}\, dx = -\ln \frac{1}{a} = \ln a.$$

This is negative for $a < 1$, zero for $a = 1$, and positive for $a > 1$, and as $a \to 0$ the density concentrates toward a point and $h \to -\infty$. A discrete entropy cannot be negative (Proposition 3.1), so h measures something different. It is not the uncertainty in X. What h measures is spread against a reference of width 1. The uniform on $[0, 1]$ sits at $h = 0$. A density more concentrated than that has $h < 0$, and one more spread out has $h > 0$. Through $H(X_\Delta) = h + \ln(1/\Delta) + o(1)$, a negative h says the variable costs less to specify at resolution Δ than the unit interval does.

The running example for the continuous case is the Gaussian, the bell curve. It is the distinguished density on the line. Among all densities of a given spread it has the most entropy, and it is the shape a diffusing density takes. The sections that follow make both precise.

Definition 5.2. A random variable X is *normal*, or *Gaussian*, with mean μ and variance $\sigma^2 > 0$, written $X \sim N(\mu, \sigma^2)$, if it has density

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

The graph is a single symmetric hump, peaked at $x = \mu$ and tapering to zero on both sides, and the two parameters fix its position and width. Changing μ slides the hump along the line without altering its shape, so μ is the center, and it is the mean $\mathbb{E}[X]$. The parameter σ sets the horizontal scale. The density depends on x only through $(x-\mu)/\sigma$, the distance from the center in units of σ, so doubling σ stretches the bell sideways by a factor of two and lowers it to keep the total area 1. A small σ is a tall narrow spike, a large one a low wide spread. Inside the exponential the distance from the center enters squared, so the density falls off quickly in the tails. The constant in front, $(2\pi\sigma^2)^{-1/2}$, is whatever makes the area 1, since $\int \exp\big(-(x-\mu)^2/(2\sigma^2)\big)\,dx = \sqrt{2\pi\sigma^2}$. Its differential entropy has a closed form.

Proposition 5.3. *For $X \sim N(\mu, \sigma^2)$,*

$$h(X) = \tfrac{1}{2}\ln(2\pi e\sigma^2).$$

Proof. Differential entropy is the average surprise, $h(X) = \mathbb{E}[-\ln p(X)]$, so we first find the surprise $-\ln p(x)$ of a single value and then average it.

Take the logarithm of the density. It is a sum of two logarithms, one for each factor. The first factor is the power $(2\pi\sigma^2)^{-1/2}$, and the logarithm brings its exponent down in front. The second is an exponential, which ln undoes, leaving the exponent itself:

$$\ln p(x) = -\tfrac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}.$$

The surprise is the negative of this,

$$-\ln p(x) = \tfrac{1}{2}\ln(2\pi\sigma^2) + \frac{(x-\mu)^2}{2\sigma^2},$$

a constant plus a term that grows with the squared distance of x from the mean, so values far out in the tails are more surprising.

Now average over X. The expectation of a sum is the sum of the expectations, so it goes through term by term:

$$h(X) = \mathbb{E}[-\ln p(X)] = \mathbb{E}\big[\tfrac{1}{2}\ln(2\pi\sigma^2)\big] + \mathbb{E}\left[\frac{(X-\mu)^2}{2\sigma^2}\right].$$

The first term is a constant, and the expectation of a constant is itself. In the second, the factor $1/(2\sigma^2)$ is a constant and slides outside the expectation:

$$h(X) = \tfrac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\,\mathbb{E}\big[(X-\mu)^2\big].$$

The expectation $\mathbb{E}[(X-\mu)^2]$ is the *variance* of X, the average squared distance of a draw from its mean μ. The squared distance and not the signed one, since $X - \mu$ averages to zero, draws landing above and below the mean in equal measure, so the square is what registers spread. We compute it for the normal. Writing the expectation as a density-weighted integral and substituting $u = x - \mu$,

$$\mathbb{E}[(X-\mu)^2] = \int (x-\mu)^2 p(x)\,dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} u^2\, e^{-u^2/(2\sigma^2)}\,du.$$

The derivative of $e^{-u^2/(2\sigma^2)}$ is $-\frac{u}{\sigma^2}\,e^{-u^2/(2\sigma^2)}$, so $u\,e^{-u^2/(2\sigma^2)} = -\sigma^2\,\frac{d}{du}\,e^{-u^2/(2\sigma^2)}$. Integrating by parts,

$$\int_{-\infty}^{\infty} u^2\,e^{-u^2/(2\sigma^2)}\,du = \Big[-\sigma^2 u\,e^{-u^2/(2\sigma^2)}\Big]_{-\infty}^{\infty} + \sigma^2\int_{-\infty}^{\infty} e^{-u^2/(2\sigma^2)}\,du.$$

The boundary term is 0, since $e^{-u^2/(2\sigma^2)}$ dies off faster than u grows, and the remaining integral is $\sqrt{2\pi\sigma^2}$, the same fact that makes p integrate to 1. So the right side is $\sigma^2\sqrt{2\pi\sigma^2}$, and

$$\mathbb{E}[(X-\mu)^2] = \frac{1}{\sqrt{2\pi\sigma^2}}\cdot\sigma^2\sqrt{2\pi\sigma^2} = \sigma^2.$$

The variance is σ^2, exactly the parameter in the name $N(\mu,\sigma^2)$. The second term is then

$$\frac{1}{2\sigma^2}\cdot\sigma^2 = \frac{1}{2}.$$

Therefore

$$h(X) = \tfrac{1}{2}\ln(2\pi\sigma^2) + \tfrac{1}{2}.$$

To collect this under one logarithm, write $\frac{1}{2} = \frac{1}{2}\ln e$, using $\ln e = 1$. Both terms now have the form $\frac{1}{2}\ln(\cdot)$, and $\ln a + \ln b = \ln(ab)$ combines them:

$$h(X) = \tfrac{1}{2}\ln(2\pi\sigma^2) + \tfrac{1}{2}\ln e = \tfrac{1}{2}\ln(2\pi e\sigma^2). \qquad \square$$

The mean μ does not appear, since shifting a density along the line does not change its spread. Only the variance enters, through $\ln\sigma^2$. The entropy is zero when $2\pi e\sigma^2 = 1$, at $\sigma = 1/\sqrt{2\pi e} \approx 0.242$, and negative for any narrower Gaussian. Figure 5 plots h against σ.

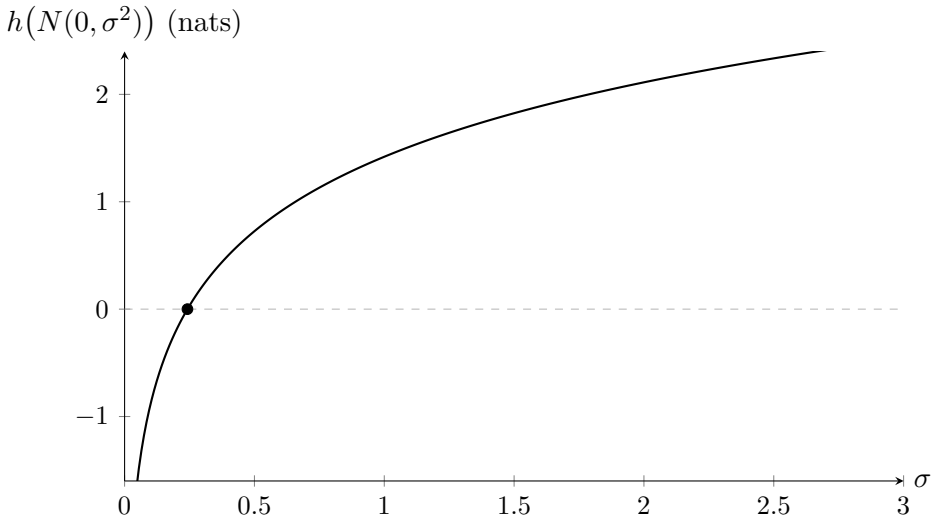

Figure 5: Differential entropy of $N(0,\sigma^2)$ against σ. The curve is $\frac{1}{2}\ln(2\pi e\sigma^2) = \ln\sigma + \frac{1}{2}\ln(2\pi e)$, which crosses zero at $\sigma = 1/\sqrt{2\pi e} \approx 0.242$ (marked) and is negative for narrower Gaussians.

The coordinate dependence is the second property. Suppose $Y = g(X)$ for a monotone differentiable g. Its density follows from conserving probability. The event $X \in [x, x+dx]$ is the same event as $Y \in [y, y+dy]$ at $y = g(x)$, so the two carry equal probability, $p_Y(y)\,|dy| = p_X(x)\,|dx|$. Since $|dy| = |g'(x)|\,|dx|$, dividing gives $p_Y(y) = p_X(x)/|g'(x)|$. The density divides by $|g'|$ because g

stretches the line by that factor near x, and the same probability over more length is a lower density. Putting this into $h(Y) = -\int p_Y \ln p_Y \, dy$ and substituting $dy = |g'(x)|\, dx$ to return the integral to x,

$$h(Y) = -\int \frac{p_X(x)}{|g'(x)|} \ln \frac{p_X(x)}{|g'(x)|} \, |g'(x)| \, dx = -\int p_X(x) \ln \frac{p_X(x)}{|g'(x)|} \, dx,$$

where the outer $|g'(x)|$ cancels the one dividing the density. Splitting the logarithm,

$$h(Y) = -\int p_X(x) \ln p_X(x) \, dx + \int p_X(x) \ln |g'(x)| \, dx = h(X) + \mathbb{E}\big[\ln |g'(X)|\big].$$

For a linear change $Y = aX$ the derivative is constant, and this reads

$$h(aX) = h(X) + \ln |a|.$$

Stretching the axis by a factor a adds $\ln |a|$ to the entropy. The discrete case has no analogue. Relabeling the outcomes of a discrete variable leaves H unchanged, since H depends only on the probabilities p_i and not on the values x_i. Differential entropy depends on the values, for the reason given in Section 5.1. Both properties follow from h being measured against a reference fixed on the line. Section 6 identifies that reference and shows why differences of entropies are free of both.

5.3 The Gaussian and diffusion

Entropy so far has been a fixed number for a fixed density. But densities evolve. Ink spreads through water, noise accumulates in a signal, and in both cases the spread grows with time and the entropy with it. The static definition cannot see this. It gives entropy at an instant, not the rate at which a spreading density produces it. That rate is what ties entropy to a differential equation, and Section 6 is pointed at it.

The Gaussian is where to start, its entropy already known and its spreading simple to follow. Since that entropy depends on the variance alone (Proposition 5.3), with the mean playing no part, we center the bell at the origin, set $\mu = 0$, and let the variance grow with time as $\sigma^2 = t$.

This turns the single density into a function of two variables,

$$p(x,t) = \frac{1}{\sqrt{2\pi t}} \exp\!\left(-\frac{x^2}{2t}\right),$$

the density $N(0,\sigma^2)$ from before with σ^2 replaced by t. The two arguments play different roles. Fix t at any value and $p(x,t)$, read as a function of x, is an ordinary density, a single bell, the $N(0,t)$ one. Let t vary and the bell changes shape, so the second argument is time and $p(x,t)$ records the whole evolution at once, the height of the curve at position x at time t. At small t the bell is a tall narrow spike at the origin, and as t grows it flattens and spreads along the line.

This spreading is not arbitrary. We reached the family by hand, taking the Gaussian and letting its variance grow as $\sigma^2 = t$, with no equation in sight. It satisfies one anyway. The same bell that gave us an entropy is, once set spreading, a solution of the heat equation. Writing that equation calls for derivatives of p in each of its two variables, and since $p(x,t)$ depends on both, a derivative has to say which variable is moving. We write $\partial p/\partial t$, or $\partial_t p$ for short, for the derivative in t with x held fixed, and $\partial p/\partial x$, or $\partial_x p$, for the derivative in x with t held fixed. The second derivative in x is $\partial^2 p/\partial x^2 = \partial_x^2 p$. These are partial derivatives, each taking one variable as live and freezing the other.

Proposition 5.4. *The density* $p(x,t) = (2\pi t)^{-1/2}\exp(-x^2/2t)$ *satisfies the heat equation*

$$\frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial^2 p}{\partial x^2},$$

the equation governing how a density spreads by diffusion.

The left side $\partial_t p$ is the rate at which the density at a fixed point x changes as time passes. The right side is built from the curvature in x. The second derivative $\partial_x^2 p$ is positive where the graph of p is concave up, in the valleys away from the peak, and negative where it is concave down, near the peak. The equation links the two. Where p is concave up the density rises, filling in, and where it is concave down it falls, drawing the peak down. The bell flattens and spreads, which is exactly diffusion.

Proof. We compute both sides of the equation from the formula for p and check that they agree. The density is a product of a power of t and an exponential, and its logarithm is easier to differentiate than the density itself,

$$\ln p = -\tfrac{1}{2}\ln(2\pi t) - \frac{x^2}{2t}.$$

A derivative of p comes from a derivative of $\ln p$ by the chain rule. Since $\partial(\ln p) = (\partial p)/p$, we have $\partial p = p\,\partial(\ln p)$, where ∂ stands for whichever of ∂_t and ∂_x we are taking, so the identity serves for both variables.

Differentiate in t, holding x fixed. The first term gives $\partial_t[-\frac{1}{2}\ln(2\pi t)] = -\frac{1}{2t}$, and the second gives $\partial_t[-x^2/(2t)] = x^2/(2t^2)$, so

$$\partial_t \ln p = -\frac{1}{2t} + \frac{x^2}{2t^2}, \qquad \partial_t p = p\left(-\frac{1}{2t} + \frac{x^2}{2t^2}\right).$$

Differentiate in x, holding t fixed. The first term has no x and drops, and the second gives $\partial_x[-x^2/(2t)] = -x/t$, so

$$\partial_x \ln p = -\frac{x}{t}, \qquad \partial_x p = -\frac{x}{t}\,p.$$

Differentiate once more, by the product rule. With $\partial_x(-x/t) = -1/t$ and $\partial_x p = -(x/t)\,p$,

$$\partial_x^2 p = \partial_x\Big(-\frac{x}{t}\,p\Big) = \left(-\frac{1}{t}\right)p + \Big(-\frac{x}{t}\Big)\Big(-\frac{x}{t}\,p\Big) = p\left(\frac{x^2}{t^2} - \frac{1}{t}\right).$$

Half of this is $p\left(\dfrac{x^2}{2t^2} - \dfrac{1}{2t}\right)$, which is $\partial_t p$. So $\partial_t p = \frac{1}{2}\,\partial_x^2 p$. □

The constant $\frac{1}{2}$ in front of $\partial_x^2 p$ is conventional. This value is the one that makes the diffusing density exactly $N(0,t)$, with variance t, and that is what lets the entropy come straight from Proposition 5.3.

Now watch the entropy change as t grows. At time t the density is $p(x,t)$, which is $N(0,t)$, a Gaussian with mean 0 and variance t. Let X_t be a random variable with this density, so $X_t \sim N(0,t)$. Proposition 5.3 gives the entropy of any $N(\mu,\sigma^2)$ as $\frac{1}{2}\ln(2\pi e\sigma^2)$, and here the mean is 0 and the variance is t, so setting $\sigma^2 = t$,

$$h(X_t) = \tfrac{1}{2}\ln(2\pi e t).$$

Split the logarithm of the product, $\ln(2\pi e t) = \ln(2\pi e) + \ln t$, to separate the constant from the part that moves with t,

$$h(X_t) = \tfrac{1}{2}\ln(2\pi e) + \tfrac{1}{2}\ln t.$$

The first term is fixed. Differentiating the second, with $\frac{d}{dt}\ln t = 1/t$,

$$\frac{d}{dt}\, h(X_t) = \frac{1}{2}\cdot\frac{1}{t} = \frac{1}{2t}.$$

The entropy grows without bound as $t \to \infty$ and falls to $-\infty$ as $t \to 0$, where the mass collapses back to a point. The rate $1/(2t)$ is positive for every $t > 0$ and decreases as t grows, so the diffusion always raises the entropy, quickly while the density is concentrated and slowly once it has spread.

The Gaussian is only one density, but the heat equation acts on any of them, carrying an arbitrary starting density $p(x, 0)$ to $p(x, t)$ as t runs. The same question can be asked of any such family: does its entropy rise at a rate one can write down, and what is that rate?

6 Relative Entropy and the Heat Equation

Differential entropy came with two defects. It can be negative, and it changes when the coordinate changes, because it measures a single density against a fixed reference on the line. Both defects come from that reference, and both vanish once the object of study is not one density but a comparison of two. Relative entropy, the gap between two densities, carries over from the discrete case with nothing added or lost. Its local form, at the end of this section, is what ties entropy to the heat equation.

6.1 Relative entropy in the continuous case

For two densities p and q on the real line, the relative entropy of p from q is the continuous form of the definition from Section 4, the sum now an integral.

Definition 6.1. The *relative entropy* of a density p from a density q is

$$D(p \,\|\, q) = \int_{-\infty}^{\infty} p(x)\, \ln \frac{p(x)}{q(x)}\, dx.$$

Unlike differential entropy, this quantity survives the passage from sums to integrals intact. Discretize both densities at width Δ. In each bin the mean value theorem gives one point ξ_i for the first density and another, ξ_i', for the second; the two need not coincide. The bin then carries probability exactly $p(\xi_i)\Delta$ under the first density and $q(\xi_i')\Delta$ under the second, and the discrete relative entropy is

$$\sum_i p(\xi_i)\Delta\, \ln \frac{p(\xi_i)\Delta}{q(\xi_i')\Delta} = \sum_i p(\xi_i)\Delta\, \ln \frac{p(\xi_i)}{q(\xi_i')},$$

since the Δ inside the logarithm cancels before any limit is taken. There is no divergent $\ln(1/\Delta)$ to remove, and as $\Delta \to 0$ the sum tends to $\int p \ln(p/q)$, under the same kind of hypotheses as in Section 5.1, with both densities continuous and q positive wherever p is; the points ξ_i and ξ_i' sit in the same shrinking bin, and continuity is what makes the mismatch between them vanish in the limit. The reason is the one that made differential entropy misbehave. There the $\ln(1/\Delta)$ came from a single density with no reference to cancel it. Compared at the same resolution, two densities cancel the resolution against itself.

The same cancellation makes D independent of the coordinate. Under a change of variable $y = g(x)$ with g monotone and differentiable, both densities pick up the same Jacobian, $p_Y(y) =$

$p_X(x)/|g'(x)|$ and $q_Y(y) = q_X(x)/|g'(x)|$, so their ratio $p_Y/q_Y = p_X/q_X$ is unchanged. Writing the divergence in the new coordinate and substituting $p_Y = p_X/|g'|$ together with $dy = |g'|\,dx$,

$$\int p_Y(y) \ln \frac{p_Y(y)}{q_Y(y)}\,dy = \int \frac{p_X(x)}{|g'(x)|} \ln \frac{p_X(x)}{q_X(x)}\,|g'(x)|\,dx = \int p_X(x) \ln \frac{p_X(x)}{q_X(x)}\,dx,$$

the $|g'|$ from the measure cancelling the $1/|g'|$ in the density, so the value is the same in x or in y. Differential entropy, which picked up $\ln|a|$ under $x \mapsto ax$, had no second density to supply that cancellation.

Relative entropy is never negative, by the argument used in the discrete case.

Proposition 6.2. *For densities p and q, $D(p\,\|\,q) \geq 0$, with equality if and only if $p = q$.*

Proof. The inequality $\ln t \leq t - 1$ holds for every $t > 0$, with equality only at $t = 1$. Apply it with $t = q(x)/p(x)$ wherever $p(x) > 0$, giving $\ln(q/p) \leq q/p - 1$. Multiply by $p(x) \geq 0$ and integrate over the region where $p > 0$,

$$\int p \,\ln \frac{q}{p}\,dx \leq \int_{p>0} (q - p)\,dx = \int_{p>0} q\,dx - 1 \leq 0,$$

using $\int_{p>0} q \leq \int q = 1$ and $\int_{p>0} p = 1$. The left side is $-D(p\,\|\,q)$, so $D(p\,\|\,q) \geq 0$. Equality needs both inequalities tight. The first forces $q/p = 1$ wherever $p > 0$, so $q = p$ on the support of p; the second forces $\int_{p>0} q = 1$, so q places no mass where $p = 0$. Together, $p = q$. □

Differential entropy is the special case of the flat reference. Put $m(x) = 1$, constant on the line, in place of q,

$$-D(p\,\|\,m) = -\int p \,\ln \frac{p}{1}\,dx = -\int p \ln p\,dx = h(X),$$

so h is the negative relative entropy of p from the flat density. The subtlety is that $m \equiv 1$ does not integrate to 1 and is not a probability density, so this is the same integral as the definition without being an instance of it. Both defects of h trace to this. Because m is not a probability density, the divergence from it is not forced to be non-negative, and $h = -D(p\,\|\,m)$ is free to take either sign. Because flatness is not a coordinate-free notion, h moves with the coordinate. Under $x \mapsto g(x)$ the reference $m \equiv 1$ acquires the Jacobian like any density and is no longer constant, so a density that looks flat in x looks tilted in y, and the mismatch is exactly the $\ln|g'|$ of Section 5.2. A relative entropy between two genuine densities calls on no flat reference and has neither defect. Differences of entropies, built from such comparisons, are free of both.

One consequence of non-negativity explains why the Gaussian has been the running example.

Theorem 6.3. *Among all densities on $\mathbb{R}$ with mean 0 and variance σ^2, the Gaussian $N(0, \sigma^2)$ has the greatest differential entropy, and no other density attains it.*

Proof. Let q be any density with mean 0 and variance σ^2, and let $p = N(0, \sigma^2)$, the Gaussian of the same variance. Proposition 6.2 gives $D(q\,\|\,p) \geq 0$. By the definition of relative entropy and the split $\ln(q/p) = \ln q - \ln p$,

$$D(q\,\|\,p) = \int q \ln \frac{q}{p}\,dx = \int q \ln q\,dx - \int q \ln p\,dx = -h(q) - \int q \ln p,$$

since $\int q \ln q\,dx = -h(q)$ is the differential entropy of q with a minus sign. The logarithm of the Gaussian is quadratic,

$$\ln p(x) = -\tfrac{1}{2}\ln(2\pi\sigma^2) - \frac{x^2}{2\sigma^2},$$

so the second integral $\int q \ln p$ depends on q only through $\int q\,dx = 1$ and $\int x^2 q\,dx = \sigma^2$, the variance at mean 0,

$$\begin{aligned}\int q \ln p\,dx &= \int q\left(-\tfrac{1}{2}\ln(2\pi\sigma^2) - \frac{x^2}{2\sigma^2}\right)dx \\ &= -\tfrac{1}{2}\ln(2\pi\sigma^2)\int q\,dx - \frac{1}{2\sigma^2}\int x^2 q\,dx = -\tfrac{1}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2}.\end{aligned}$$

Both quantities are shared by every density in the class, p among them, so the same computation with p in place of q returns the identical value. Hence $\int q \ln p = \int p \ln p = -h(p)$, and

$$D(q \,\|\, p) = -h(q) + h(p) = h(p) - h(q) \geq 0.$$

Therefore $h(q) \leq h(p) = \frac{1}{2}\ln(2\pi e\sigma^2)$, with equality only when $D(q \,\|\, p) = 0$, that is $q = p$. □

Remark 6.4. Maximizing entropy under known constraints is a principle that runs well past this example. Jaynes [Jay57] read the maximum-entropy distribution as the least presumptuous choice consistent with what is known, and from that principle recovered the distributions of statistical mechanics. The Gaussian is the case where the constraint is a fixed variance.

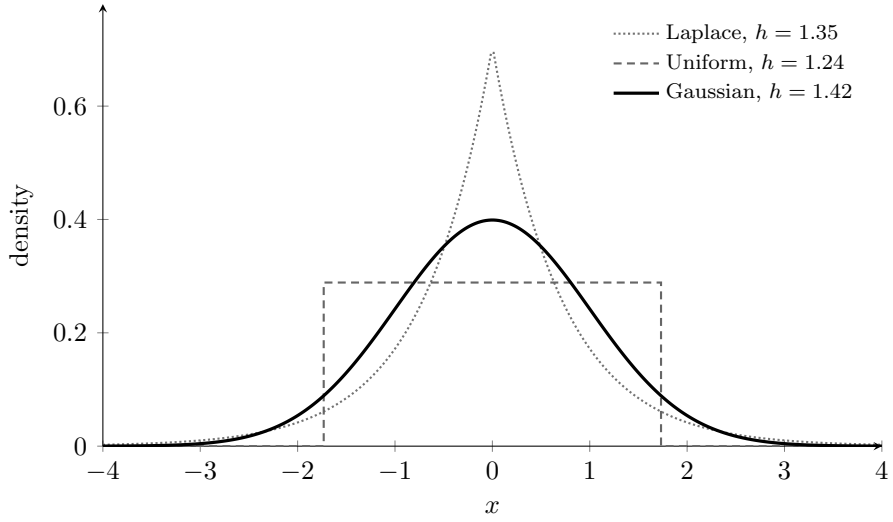

Figure 6: Three densities with the same mean 0 and variance 1: a Gaussian, a Laplace density (more peaked), and a uniform (flat, on $[-\sqrt{3}, \sqrt{3}]$). They differ in shape but share the same spread, and among all such densities the Gaussian has the greatest differential entropy, $h = \frac{1}{2}\ln(2\pi e) \approx 1.42$ nats, against 1.35 for the Laplace and 1.24 for the uniform.

The Gaussian is the maximum-entropy density at fixed variance, and, from Section 5.3, it is also the heat kernel, the density a point of mass spreads into as it diffuses. One density stands at both ends, the most uncertain one for a given spread and the shape diffusion produces. Section 6.3 makes the link quantitative, computing the rate at which diffusion raises a density's entropy.

6.2 Mutual information in the continuous case

Mutual information carries over to densities in the same way. For continuous X and Y with joint density $p(x, y)$ and marginals $p(x)$ and $p(y)$, it is the relative entropy of the joint density from the

product of the marginals,

$$I(X;Y) = D\big(p(x,y) \,\|\, p(x)p(y)\big) = \iint p(x,y)\, \ln \frac{p(x,y)}{p(x)\,p(y)}\, dx\, dy.$$

Being a relative entropy, it has both of the properties established in Section 6.1. It is non-negative, and zero exactly when $p(x,y) = p(x)p(y)$, which is independence, by Proposition 6.2 applied to the joint density and the product of the marginals. And it is coordinate-independent, unchanged when X or Y is reparametrized, because relative entropy is.

The coordinate-independence is easy to lose sight of when I is written through entropies. The identity of Section 4 carries over by the same algebra (we state it through joint and marginal entropies, sidestepping the conditional density $p(y \mid x)$ that a direct analogue of Section 3 would require),

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

and each of the three entropies on the right depends on the coordinate. Their combination does not. Rescaling X by a adds $\ln|a|$ to both $h(X)$ and $h(X,Y)$, where the two cancel, leaving I fixed. The dependence that made differential entropy awkward drops out of any quantity built as a relative entropy, this one included.

One example carries most of the content, a pair of correlated Gaussians. Let (X,Y) be jointly normal with mean zero and correlation ρ. Rescaling to unit variance changes neither I nor ρ, so take $\sigma_X = \sigma_Y = 1$, and the joint density is

$$p(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\, \exp\!\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right),$$

the two-dimensional Gaussian, the exponential of a negative quadratic in x and y normalized to integrate to 1, with $\rho = \mathbb{E}[XY]$ the correlation and $|\rho| < 1$. The marginals are each $N(0,1)$, so $h(X) = h(Y) = \frac{1}{2}\ln(2\pi e)$ by Proposition 5.3, and only the joint entropy is left to find.

The surprise of the joint density is

$$-\ln p(x,y) = \ln\big(2\pi\sqrt{1-\rho^2}\,\big) + \frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)},$$

and averaging over (X,Y) keeps the constant and replaces the quadratic by its expectation,

$$h(X,Y) = \ln\big(2\pi\sqrt{1-\rho^2}\,\big) + \frac{\mathbb{E}[X^2] - 2\rho\,\mathbb{E}[XY] + \mathbb{E}[Y^2]}{2(1-\rho^2)}.$$

With unit variances and correlation ρ, $\mathbb{E}[X^2] = \mathbb{E}[Y^2] = 1$ and $\mathbb{E}[XY] = \rho$, so the numerator is $1 - 2\rho^2 + 1 = 2(1-\rho^2)$ and the fraction equals 1. Hence

$$h(X,Y) = \ln\big(2\pi\sqrt{1-\rho^2}\,\big) + 1 = \ln(2\pi e) + \tfrac{1}{2}\ln(1-\rho^2),$$

using $\ln\sqrt{1-\rho^2} = \frac{1}{2}\ln(1-\rho^2)$ and $\ln(2\pi) + 1 = \ln(2\pi e)$. The three entropies combine into

$$I(X;Y) = \tfrac{1}{2}\ln(2\pi e) + \tfrac{1}{2}\ln(2\pi e) - \ln(2\pi e) - \tfrac{1}{2}\ln(1-\rho^2) = -\tfrac{1}{2}\ln(1-\rho^2).$$

At $\rho = 0$ the variables are independent and $I = 0$. As $\rho \to \pm 1$ each variable comes to determine the other and $I \to \infty$. Between these, the information one variable carries about the other depends only on the correlation and not on the scales, as an invariant quantity should. The entropies it was assembled from shift with the coordinate and can be negative. The dependence itself is a single finite number, free of both.

6.3 Fisher information and de Bruijn's identity

Section 4 named the Fisher information but did not define it. The claim there was that the relative entropy between a density and a slightly shifted copy of it shrinks like the square of the shift, with the Fisher information as the coefficient. This section makes that precise, and then shows that the same quantity fixes the rate at which a diffusing density gains entropy, the question left open at the end of Section 5.

Fix a density p and compare it to $p_\delta(x) = p(x-\delta)$, the same density slid to the right by δ. Their relative entropy is

$$D(p \,\|\, p_\delta) = \int p(x)\, \ln \frac{p(x)}{p(x-\delta)}\, dx = \int p(x) \big[\ln p(x) - \ln p(x-\delta) \big]\, dx.$$

The tool is the Taylor expansion: near a point a, a smooth function is its value there, corrected by its slope times the step, then by half its second derivative times the step squared, and so on in higher powers of the step,

$$f(a+h) = f(a) + f'(a)\, h + \tfrac{1}{2} f''(a)\, h^2 + O(h^3),$$

each higher power of the step carrying the matching derivative divided by a factorial, so that the two sides agree ever more closely as $h \to 0$.[2] Apply it to $f = \ln p$ at the point $a = x$ with step $h = -\delta$, since $p(x-\delta) = p\big(x + (-\delta)\big)$. The linear term is $(\ln p)'(x)\,(-\delta)$ and the quadratic term is $\frac{1}{2}(\ln p)''(x)\,(-\delta)^2$, the minus surviving in the first and squared away in the second,

$$\ln p(x-\delta) = \ln p(x) - \delta\, (\ln p)'(x) + \tfrac{\delta^2}{2} (\ln p)''(x) + O(\delta^3).$$

Subtracting this from $\ln p(x)$ leaves $\ln p(x) - \ln p(x-\delta) = \delta\, (\ln p)' - \frac{\delta^2}{2}(\ln p)'' + O(\delta^3)$, so, assuming p is smooth enough and its derivatives decay fast enough at infinity for the remainder to integrate to $O(\delta^3)$,

$$D(p \,\|\, p_\delta) = \delta \int p\, (\ln p)'\, dx - \frac{\delta^2}{2} \int p\, (\ln p)''\, dx + O(\delta^3).$$

The linear term vanishes. Its integral is $\int p\, (\ln p)'\, dx = \int p'\, dx = 0$, since $(\ln p)' = p'/p$ and p vanishes at both ends. It must vanish on other grounds as well, since $D(p \,\|\, p_\delta) \geq 0$ with equality at $\delta = 0$ makes $\delta = 0$ a minimum, where the first derivative is zero. For the quadratic term, $(\ln p)'' = (p'/p)' = p''/p - (p'/p)^2$, so

$$\int p\, (\ln p)''\, dx = \int p''\, dx - \int \frac{(p')^2}{p}\, dx = -\int \frac{(p')^2}{p}\, dx,$$

using $\int p''\, dx = 0$. Writing J for the remaining integral,

$$D(p \,\|\, p_\delta) = \tfrac{1}{2}\, J\, \delta^2 + O(\delta^3), \qquad J = \int \frac{(p')^2}{p}\, dx = \mathbb{E}\big[\big((\ln p)'(X) \big)^2 \big].$$

This J is the *Fisher information* of p, the quantity Fisher [Fis22] built statistical estimation around and Rao [Rao45] tied to the best precision an estimator can reach. It is the curvature of the relative

[2] The coefficients are forced by matching. For a polynomial $c_0 + c_1 h + c_2 h^2 + \cdots$ to share the value and every derivative of f at $h = 0$, differentiating it n times and setting $h = 0$ leaves $n!\, c_n = f^{(n)}(a)$, since differentiating h^n down to a constant produces the factor $n! = n(n-1)\cdots 1$; hence $c_n = f^{(n)}(a)/n!$. Higher powers of a small step shrink fast, so cutting the expansion off leaves an error the size of the first dropped term, written $O(h^3)$ here since the expansion stops after the h^2 term.

entropy at its minimum, the coefficient Section 4 promised, and it measures how sharply p is peaked. A concentrated density changes quickly under a small shift and has large J; a flat, spread-out one changes slowly and has small J.

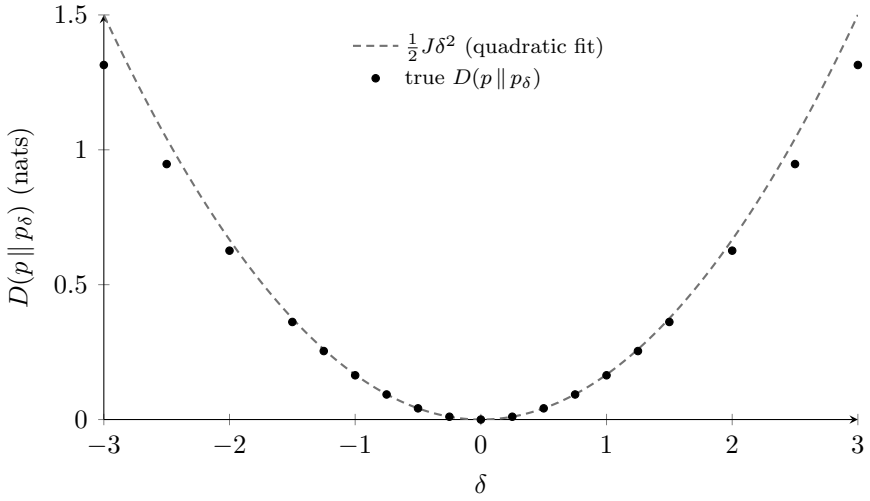

Figure 7: Relative entropy between a density p (here a standard logistic) and its shift $p_\delta(x) = p(x-\delta)$, against δ. Near $\delta = 0$ the curve is a parabola $\frac{1}{2}J\delta^2$ whose curvature is the Fisher information $J = \frac{1}{3}$; the fit is exact at the bottom and loosens as the shift grows.

The same J reappears under diffusion. Add to X independent Gaussian noise that grows with time, $X_t = X + \sqrt{t}\,Z$ with $Z \sim N(0,1)$. The density of X_t is the convolution of the density of X with the $N(0,t)$ kernel φ_t,[3]

$$p(x,t) = \int p(\xi,0)\,\varphi_t(x-\xi)\,d\xi, \qquad \varphi_t(u) = \frac{1}{\sqrt{2\pi t}}\,e^{-u^2/2t}.$$

The kernel φ_t solves the heat equation (Proposition 5.4), and since $p(\xi,0)$ carries no x or t, the derivatives ∂_t and ∂_x^2 act only on φ_t inside the integral. So $p(x,t)$ solves it too,

$$\partial_t p = p(\cdot,0) * \partial_t \varphi_t = p(\cdot,0) * \tfrac{1}{2}\partial_x^2 \varphi_t = \tfrac{1}{2}\partial_x^2 p.$$

This is a density spreading by diffusion, as in Section 5.3, but now from an arbitrary starting density rather than a point.

Follow its entropy. Differentiating $h(X_t) = -\int p \ln p\,dx$ in t,

$$\frac{d}{dt}h(X_t) = -\int \partial_t(p\ln p)\,dx = -\int (\partial_t p)\ln p\,dx - \int \partial_t p\,dx.$$

The last integral is $\int \partial_t p\,dx = \frac{d}{dt}\int p\,dx = 0$, since p remains a density. Substituting the heat equation into the first,

$$\frac{d}{dt}h(X_t) = -\tfrac{1}{2}\int (\partial_x^2 p)\,\ln p\,dx,$$

[3]When two independent variables are added, the density of the sum is the convolution of their densities. To land on $X_t = x$, the two pieces must split as $X = \xi$ and $\sqrt{t}\,Z = x - \xi$ for some ξ; the pieces are independent, so that split carries density $p(\xi,0)\,\varphi_t(x-\xi)$, and summing over all splits gives $p(x,t) = \int p(\xi,0)\,\varphi_t(x-\xi)\,d\xi$.

and integrating by parts, with the boundary term $[(\partial_x p)\ln p]$ vanishing at both ends,

$$\int (\partial_x^2 p)\,\ln p\,dx = -\int (\partial_x p)\,\frac{\partial_x p}{p}\,dx = -\int \frac{(\partial_x p)^2}{p}\,dx.$$

Therefore

$$\frac{d}{dt}h(X_t) = \tfrac{1}{2}\int \frac{(\partial_x p)^2}{p}\,dx = \tfrac{1}{2}\,J(X_t),$$

the same Fisher information as before, now of the density $p(\cdot,t)$. This is de Bruijn's identity, first published by Stam [Sta59], who credited it to de Bruijn; Cover and Thomas [CT06] give the modern treatment. The steps assume p is smooth and falls off fast enough at infinity for the boundary term to vanish and the t-derivative to pass under the integral.

The Gaussian case recovers Section 5.3. Starting from a point at the origin, $X_t = \sqrt{t}\,Z$ has density $N(0,t)$, and $\partial_x p = -(x/t)\,p$ gives $(\partial_x p)^2/p = (x^2/t^2)\,p$, so

$$J = \int \frac{x^2}{t^2}\,p\,dx = \frac{1}{t^2}\,\mathbb{E}[X_t^2] = \frac{1}{t^2}\cdot t = \frac{1}{t},$$

and $\frac{d}{dt}h = \frac{1}{2t}$, the rate found there directly. Beyond this one case, $J \geq 0$ for every density, being the integral of a square against a positive weight, so $\frac{d}{dt}h(X_t) \geq 0$. Entropy never decreases under diffusion. The heat equation spreads a density and raises its entropy, at a rate equal to half the density's Fisher information.

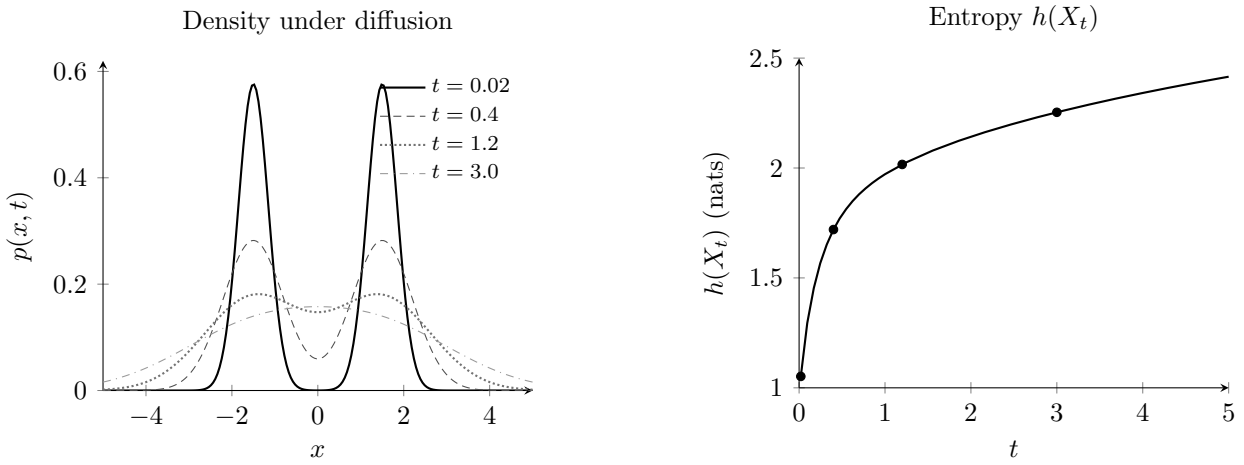

Figure 8: Left: an initial density with two bumps, evolved by the heat equation $\partial_t p = \frac{1}{2}\partial_x^2 p$. As t grows the bumps spread and merge into a single hump. Right: the differential entropy $h(X_t)$ of the same density, rising at every instant; the four marked times are the curves on the left. De Bruijn's identity says the slope equals $\frac{1}{2}J(X_t)$.

Entropy began in Section 2 as a finite sum $-\sum_i p_i \ln p_i$ over a short list of outcomes. On the real line it became an integral, less well-behaved than the sum, and the quantity that carried the good behavior across was relative entropy, the comparison of two densities. Its local form is the Fisher information, and de Bruijn's identity makes that the rate at which a diffusing density gains entropy. The equation behind the diffusion is the heat equation, the relation named in the title.

That relative entropy locally resembles a squared distance, with the Fisher information as coefficient, is not an accident of the Taylor expansion. It is the opening fact of information

geometry [Ama16, Che82], which treats families of probability distributions as curved spaces, with the Fisher information as the local ruler for distance.

References

[Ama16] S. Amari, *Information Geometry and Its Applications*. Tokyo: Springer, 2016.

[Che82] N. N. Chentsov, *Statistical Decision Rules and Optimal Inference*, Translations of Mathematical Monographs 53. Providence, RI: American Mathematical Society, 1982.

[CT06] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006.

[Fis22] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society A*, vol. 222, pp. 309–368, 1922.

[Jay57] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, no. 4, pp. 620–630, 1957.

[Khi57] A. Ya. Khinchin, *Mathematical Foundations of Information Theory*. New York: Dover, 1957.

[KL51] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[Rao45] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, vol. 37, pp. 81–91, 1945.

[Ren61] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, Berkeley, CA: Univ. of California Press, 1961, pp. 547–561.

[Sha48] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[Sha51] C. E. Shannon, "Prediction and entropy of printed English," *Bell System Technical Journal*, vol. 30, no. 1, pp. 50–64, 1951.

[Sta59] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Information and Control*, vol. 2, no. 2, pp. 101–112, 1959.