Entropy and mutual information

Last modified: 2026-11-12 01:15:28 (PST)

This page builds on the definitions of entropy, cross-entropy and Kullback–Leibler divergence in *Probability for Data Science*.

Surprisal

Definition 0.1 (Surprisal)**.** The **surprisal** of a value x of a discrete random variable X with $\mathbb{P}(X = x) > 0$ is

$$s(x) \stackrel{\text{def}}{=} -\log \mathbb{P}(X = x).$$

With base-2 logarithms, surprisal is also called the **Shannon information content** of the value x (MacKay 2003, sec. 2.4).

Definition 0.2 (Bit (shannon))**.** A **bit**, also called a **shannon** (symbol Sh) (International Electrotechnical Commission 2008), is the unit of information when the logarithm in surprisal, and in every quantity built from it, has base 2.
One bit is the surprisal of a value with probability $1/2$: for a value x of a discrete random variable X,

$$\begin{aligned} s(x) = 1 \text{ bit} &\iff -\log_2 \mathbb{P}(X = x) = 1 \\ &\iff \mathbb{P}(X = x) = 2^{-1} = \tfrac{1}{2}. \end{aligned}$$

In words, a bit is the surprisal of the outcome of one trial whose two outcomes are equally likely, such as one toss of a fair coin.

Example 0.1 (Counting bits with fair coins)**.** One toss of a fair coin has 2 equally likely outcomes, and one binary digit is enough to record which one happened. Two tosses have 4 outcomes, and two binary digits record them. In general, k tosses have $n = 2^k$ equally

likely outcomes, and $k = \log_2 n$ binary digits record them.
Each of those n outcomes has probability $p = 1/n$, so

$$\begin{aligned} k &= \log_2 n \\ &= \log_2 \frac{1}{p} \\ &= -\log_2 p, \end{aligned}$$

which is the surprisal of any one of them, in bits.
Ten tosses that all land heads have probability 2^{-10}, so

$$\begin{aligned} s(\text{ten heads}) &= -\log_2 2^{-10} \\ &= 10 \text{ bits}. \end{aligned}$$

Theorem 0.1 (Surprisal as a logarithm to base one half)**.** *For a value x of a discrete random variable X with $\mathbb{P}(X = x) > 0$, the surprisal in bits is*

$$s(x) = \log_{1/2} \mathbb{P}(X = x).$$

Proof

Proof. Write $p = \mathbb{P}(X = x)$. The change of base formula gives

$$\log_b y = \frac{\log\{y\}}{\log\{b\}}$$

for $y > 0$ and $b > 0$ with $b \neq 1$. With $b = 1/2$ and $y = p$,

$$\begin{aligned} \log_{1/2} p &= \frac{\log\{p\}}{\log\{1/2\}} \\ &= \frac{\log\{p\}}{-\log\{2\}} \\ &= -\frac{\log\{p\}}{\log\{2\}} \\ &= -\log_2 p \\ &= s(x). \end{aligned}$$

The step from the third to the fourth line is the change of base formula with $b = 2$. □

Example 0.2 (Halvings of certainty for a loaded coin)**.**
A coin lands heads with probability 0.2. The chunk computes the surprisal of heads in two ways: as $-\log_2 0.2$, and as $\log_{1/2} 0.2$.

```
p_heads <- 0.2
from_base_two <- -log(x = p_heads, base = 2)
from_base_half <- log(x = p_heads, base = 0.5)
c(from_base_two, from_base_half)
#> [1] 2.321928 2.321928
```

Both give 2.322 bits (Theorem 0.1). Halving 1 that many times gives $0.5^{2.322} = 0.2$, the probability of heads.

Remark 0.1 (Surprisal as halvings of certainty). Start from certainty, probability 1. Each halving multiplies the probability by 1/2. After s halvings the probability is $(1/2)^s$, and solving $(1/2)^s = p$ for s gives $s = \log_{1/2} p$. This count is only motivation. Theorem 0.1 proves that this s is the surprisal in bits. So the surprisal of a value is the number of times certainty must be halved to reach the probability of that value. When p is not $(1/2)^k$ for a whole number k, the number of halvings is not a whole number.

Definition 0.3 (Nat)**.** A **nat** is the unit of information when the logarithm in surprisal, and in every quantity built from it, is the natural logarithm.
One nat is the surprisal of a value with probability 1/e: for a value x of a discrete random variable X,

$$\begin{aligned} s(x) = 1 \text{ nat} &\iff -\ln \mathbb{P}(X = x) = 1 \\ &\iff \mathbb{P}(X = x) = \mathrm{e}^{-1} = 1/\mathrm{e} \approx 0.368. \end{aligned}$$

Example 0.3 (One fair coin toss in nats)**.** By Example 0.1 with $k = 1$, one toss of a fair coin landing heads has a surprisal of 1 bit. In nats, the same outcome has a surprisal of

$$\begin{aligned} s(\text{heads}) &= -\ln \tfrac{1}{2} \\ &= \ln 2 \\ &\approx 0.693 \text{ nats}. \end{aligned}$$

Remark 0.2 (Converting between bits and nats). For any probability $p \in (0, 1]$, the change-of-base rule for logarithms gives

$$
\begin{aligned}
-\log_2 p &= \frac{-\ln p}{\ln 2} \\
&\approx 1.443 \times (-\ln p).
\end{aligned}
$$

So a value with a surprisal of 1 nat has a surprisal of $1/\ln 2 \approx 1.443$ bits, and the $\ln 2 \approx 0.693$ nats of Example 0.3 are exactly $\ln 2/\ln 2 = 1$ bit.
A bit as a unit of information is not the same thing as a bit as a binary digit. The name shannon refers only to the unit of information, so it avoids this ambiguity. A binary digit whose two values are equally likely has a surprisal of exactly 1 bit, whichever value it takes. A binary digit that is 0 nine times in ten does not: the value 0 has a surprisal of $-\log_2 0.9 \approx 0.152$ bits, and the value 1 has a surprisal of $-\log_2 0.1 \approx 3.322$ bits.

Remark 0.3 (Unlikely outcomes carry more information). The surprisal of a value is large when the value was unlikely, and it falls to 0 as the value becomes certain. Correctly guessing heads on one toss of a fair coin is worth 1 bit; correctly guessing ten heads in a row is worth 10 bits.

Remark 0.4 (Entropy is expected surprisal). The entropy of a discrete random variable X is the expectation of its surprisal:

$$
\begin{aligned}
H(X) &= \mathbb{E}[-\log \mathbb{P}(X)] \\
&= \mathbb{E}[s(X)].
\end{aligned}
$$

So $H(X)$ is the average number of bits (with base-2 logarithms) needed to record a value of X when each value x costs $s(x)$ bits. The source coding theorem, on a later page of these notes, makes that average precise: no uniquely decodable code does better.

Definition 0.4 (Rarity)**.** The **rarity** of a value x of a discrete random variable X is the reciprocal of its probability mass:

$$
\mathbb{r}(x) \stackrel{\text{def}}{=} \frac{1}{\mathbb{P}(X = x)}.
$$

For a continuous random variable X with density $\mathbb{p}$, the rarity of a value x is the reciprocal of the density:

$$
\mathbb{r}(x) \stackrel{\text{def}}{=} \frac{1}{\mathbb{p}(x)}.
$$

When the mass or the density is 0, the rarity is ∞.

Example 0.4 (Rarity of the outcomes of a die)**.**
Let X be the roll of a fair six-sided die. Each outcome has probability $1/6$.

```
rarity <- function(p) 1 / p
p_fair <- 1 / 6
p_loaded_six <- 1 / 2
p_loaded_other <- (1 - p_loaded_six) / 5
rarity_fair <- rarity(p_fair)
rarity_loaded_six <- rarity(p_loaded_six)
rarity_loaded_other <- rarity(p_loaded_other)
rarities <- c(
  fair = rarity_fair,
  loaded_six = rarity_loaded_six,
  loaded_other = rarity_loaded_other
)
rarities
#>         fair   loaded_six loaded_other
#>            6            2           10
```

So each face of the fair die has rarity $\mathrm{r}(x) = 6$: one roll in six.
Suppose the die is loaded so that a six has probability $1/2$. Then $\mathrm{r}(6) = 2$. If the other five faces share the rest equally, each has $\mathrm{r}(x) = 10$. The more probable an outcome, the lower its rarity.

Remark 0.5 (Source of the name). Leinster writes that, for a community whose ith species has relative abundance p_i, "$1/p_i$ is a measure of its rarity or specialness" (Leinster 2021, sec. 2.4, p. 52). The name "rarity" for the reciprocal of a probability is borrowed from there.

Theorem 0.2 (Surprisal is the logarithm of the rarity)**.** *Let X be a discrete random variable, and let x be a value of X with $\mathbb{P}(X = x) > 0$. Then the surprisal of x is the logarithm of its rarity:*

$$\mathrm{s}(x) = \log\{\mathrm{r}(x)\}.$$

Proof

Proof.

$$\begin{aligned} \mathfrak{s}(x) &= -\log\{\mathbb{P}(X = x)\} && \text{(definition of surprisal)} \\ &= \log\left\{\frac{1}{\mathbb{P}(X = x)}\right\} && \text{(logarithm of a reciprocal)} \\ &= \log\{\mathfrak{r}(x)\}. && \text{(definition of rarity)} \end{aligned}$$

□

Theorem 0.3 (Surprisal of a continuous value is the logarithm of the rarity)**.** *Let X be a continuous random variable with density $\mathfrak{p}$, and let x be a value with $\mathfrak{p}(x) > 0$. Then the surprisal of x is the logarithm of its rarity:*

$$\mathfrak{s}(x) = \log\{\mathfrak{r}(x)\}.$$

Proof

Proof.

$$\begin{aligned} \mathfrak{s}(x) &= -\log\{\mathfrak{p}(x)\} && \text{(definition of surprisal)} \\ &= \log\left\{\frac{1}{\mathfrak{p}(x)}\right\} && \text{(logarithm of a reciprocal)} \\ &= \log\{\mathfrak{r}(x)\}. && \text{(definition of rarity)} \end{aligned}$$

□

Corollary 0.1 (Surprisal of a discrete value is nonnegative)**.** *Let X be a discrete random variable, and let x be a value of X with $\mathbb{P}(X = x) > 0$. Then the rarity of x is at least 1, with equality exactly when $\mathbb{P}(X = x) = 1$, and the surprisal satisfies*

$$\mathfrak{s}(x) \geq 0.$$

Proof

Proof. Since $0 < \mathbb{P}(X = x) \leq 1$, the rarity $\mathfrak{r}(x) = 1/\mathbb{P}(X = x)$ is at least 1, with equality exactly when $\mathbb{P}(X = x) = 1$. The logarithm is increasing, and $\log\{1\} = 0$, so by Theorem 0.2, $\mathfrak{s}(x) = \log\{\mathfrak{r}(x)\} \geq 0$. □

Example 0.5 (Surprisal and rarity of a value with probability one quarter)**.** Take a value with probability $1/4$:

```
p_four <- 1 / 4
rarity_four <- 1 / p_four
surprisal_four <- log2(rarity_four)
surprisal_four
#> [1] 2
```

It has rarity 4 by definition. By Theorem 0.2, its surprisal is the base-2 logarithm of that rarity, 2 bits.

Remark 0.6 (Reading the rarity of a value). A density can exceed 1, so for a continuous random variable the rarity can be below 1, and then the surprisal is negative (Theorem 0.3). For a mass, a rarity of r reads "one in r", and the surprisal is the logarithm of r (Theorem 0.2).

Remark 0.7 (Values of the probability, the rarity and the surprisal). For a discrete random variable, the mass $\mathbb{P}(X = x)$ of a value with a mass above 0, its rarity $\mathbb{r}(x)$ and its surprisal $\mathbb{s}(x)$ take the values in the middle column. For a continuous random variable, the density $\mathbb{p}(x)$ takes the place of the mass. At a value x with a density above 0, the rarity $\mathbb{r}(x)$ is the reciprocal of $\mathbb{p}(x)$, and the surprisal $\mathbb{s}(x)$ is the surprisal of a continuous value. Their values are in the right column.

	Mass $\mathbb{P}(X = x)$	Density $\mathbb{p}(x)$
Mass or density	$(0, 1]$	$(0, \infty)$
Rarity $\mathbb{r}(x)$	$[1, \infty)$	$(0, \infty)$
Surprisal $\mathbb{s}(x)$	$[0, \infty)$	$(-\infty, \infty)$

- **Mass.** A mass of 1 gives the smallest rarity, 1, and the smallest surprisal, 0. No mass exceeds 1, the probability of a value that is certain. So no mass gives a rarity below 1 or a surprisal below 0.
- **Density.** A density can be arbitrarily large, as in the uniform example, so the rarity can fall below 1 and the surprisal can fall below 0.
- **Zero.** A mass or density of 0 gives a rarity and a surprisal of ∞.
- **Units.** A mass has no units. A density has the units of $1/x$, so its rarity has the units of x. Rescaling x by a nonzero c shifts its surprisal by $\log\{|c|\}$.

Remark 0.8 (Rarity and odds). Let x be a value of a discrete random variable X, and let A be the event $X = x$ with probability $p = \mathbb{P}(X = x)$. The odds in favour of A are $p/(1 - p)$, and the odds against A are $(1 - p)/p$ (Grinstead and Snell 1997, Definition

1.4). The rarity of x is $\mathrm{r}(x) = 1/p$, so the odds against A are $\mathrm{r}(x) - 1$.

```
p_value <- c(1 / 4, 1 / 100, 1)
rarity_value <- 1 / p_value
odds_against <- (1 - p_value) / p_value
rarity_value - odds_against
#> [1] 1 1 1
rarity_value
#> [1]   4 100   1
odds_against
#> [1]  3 99  0
```

The first printed line shows that, for each p, the rarity of the value exceeds the odds against its event by exactly 1. A value with probability 1/4 has rarity 4, and its event has odds of 3 to 1 against. The two differ by exactly 1, so they are relatively close when the rarity is large. At $p = 1$ the rarity is 1 and the odds against are 0, but the odds in favour divide by zero.

The odds compare an event with its complement, and the odds against take values in $[0, \infty)$ when p is above 0. At $p = 0$, the odds against and the rarity divide by zero, and both tend to ∞ as p decreases to 0. For a value of a discrete random variable with p above 0, the rarity takes values in $[1, \infty)$. For a continuous random variable, the event $X = x$ for a single value has probability 0, so its odds against and $1/\mathbb{P}(X = x)$ are not defined, but the rarity of the value is defined from the density. An event such as $X \leq c$ can have a probability above 0, so it can have odds.

Definition 0.5 (Perplexity)**.** The **perplexity** of a discrete random variable X is the exponential of its entropy, with the same base as the logarithm in the entropy:

$$\mathrm{PP}(X) \stackrel{\text{def}}{=} b^{\mathbb{H}(X)}.$$

With base-2 logarithms, $\mathrm{PP}(X) = 2^{\mathbb{H}(X)}$ with $\mathbb{H}(X)$ in bits. With natural logarithms, $\mathrm{PP}(X) = \mathrm{e}^{\mathbb{H}(X)}$ with $\mathbb{H}(X)$ in nats. The two give the same value.

Theorem 0.4 (Perplexity is the probability-weighted geometric mean of the rarities)**.** *For a discrete random variable X, the perplexity is a weighted geometric mean of the rarities, each raised to its probability. Take x over the values with $\mathbb{P}(X = x) > 0$. Then*

$$\mathrm{PP}(X) = \prod_x \mathrm{r}(x)^{\mathbb{P}(X=x)}.$$

Proof

Proof.

$$
\begin{aligned}
\mathrm{PP}(X) &= b^{\mathbb{H}(X)} && \text{(definition of perplexity)} \\
&= b^{\mathbb{E}[\mathbb{s}(X)]} && \text{(entropy is expected surprisal)} \\
&= b^{\sum_x \mathbb{P}(X=x) \log_b \mathbb{r}(x)} && \text{(surprisal is the logarithm of the rarity)} \\
&= \prod_x b^{\mathbb{P}(X=x) \log_b \mathbb{r}(x)} && \text{(a sum in the exponent is a product)} \\
&= \prod_x \mathbb{r}(x)^{\mathbb{P}(X=x)} && (b^{\log_b y} = y)
\end{aligned}
$$

□

Example 0.6 (Perplexity of a uniform distribution)**.**
If X is uniform on k values, then every value has probability $1/k$ and rarity k, so by Theorem 0.4 the perplexity of X is k. In general, the perplexity is the count of equally likely values that would give the same entropy as X.

Example 0.7 (Perplexity of a loaded die)**.**
Let X be the roll of the loaded die from Example 0.4, where a six has probability $1/2$ and each other face has probability $1/10$. The code computes the perplexity from the entropy in bits, and then from the rarities by Theorem 0.4.

```
p_other_faces <- rep(1 / 10, 5)
p <- c(1 / 2, p_other_faces)
surprisal <- -log2(p)
entropy <- sum(p * surprisal)
perplexity_from_entropy <- 2^entropy
rarity <- 1 / p
perplexity_from_rarities <- prod(rarity^p)
c(perplexity_from_entropy, perplexity_from_rarities)
#> [1] 4.472136 4.472136
```

Both routes give a perplexity of about 4.47. The loaded die has the same entropy as a fair die with about that many faces, fewer than the 6 faces it has.

Remark 0.9 (Source of the name). The name comes from speech recognition (Jelinek et al. 1977). Leinster defines the "diversity of order 1" as the exponential of the entropy (Leinster 2021, Definition 2.4.1). He calls a diversity index an "effective number" when

it is k for k species with the same abundance (Leinster 2021, Definition 2.4.5).

Remark 0.10 (Differential entropy differs from entropy). A density value is not a probability, so the surprisal of a continuous value and the entropy of a continuous random variable (the differential entropy) behave differently from their counterparts for a discrete random variable:

- **Sign.** A density can exceed 1, so the surprisal $\mathbb{s}(x)$ and the differential entropy can fall below 0, as in the uniform example (Cover and Thomas 2006, sec. 8.1). The entropy of a discrete random variable is never negative.
- **Units.** Rescaling changes the differential entropy: $\mathbb{H}(aX) = \mathbb{H}(X) + \log|a|$ for a constant $a \neq 0$ (Cover and Thomas 2006, Theorem 8.6.4). Measuring the same quantity in metres rather than centimetres therefore changes it. Adding a constant does not: $\mathbb{H}(X + c) = \mathbb{H}(X)$ (Cover and Thomas 2006, Theorem 8.6.3).
- **Not what the entropy approaches.** Round X down to a multiple of Δ, which gives a discrete random variable X^Δ. Then $\mathbb{H}(X^\Delta) + \log\Delta \to \mathbb{H}(X)$ as $\Delta \to 0$ (Cover and Thomas 2006, Theorem 8.3.1). The entropy $\mathbb{H}(X^\Delta)$ itself grows indefinitely as $\Delta \to 0$, so the differential entropy is not what it approaches.

Definition 0.6 (Collision probability)**.** The **collision probability** of a discrete random variable X is the chance that X and a second variable X' take the same value, where X and X' are IID:

$$\mathbb{P}(X = X').$$

Definition 0.7 (Rényi entropy of order 2 (collision entropy))**.** The **Rényi entropy of order 2**, also called the **collision entropy** (Wikipedia contributors 2026), of a discrete random variable X is the negative logarithm of the expectation of its probability, where $\mathbb{P}(X)$ is the probability mass function evaluated at the random variable X itself:

$$\mathbb{H}_2(X) = -\log\mathbb{E}[\mathbb{P}(X)].$$

Rényi introduced this case of a family of entropies (Rényi 1961).

Theorem 0.5 (The expected probability is the collision probability)**.** *Let X be a discrete random variable, and let X' be a variable such that X and X' are IID. Then*

$$\mathbb{E}[\mathbb{P}(X)] = \sum_x \mathbb{P}(X = x)^2 = \mathbb{P}(X = X').$$

Proof

Proof. The expectation of $\mathbb{P}(X)$ adds, over the values x, the value $\mathbb{P}(X = x)$ times its probability $\mathbb{P}(X = x)$:

$$\begin{aligned}\mathbb{E}[\mathbb{P}(X)] &= \sum_x \mathbb{P}(X = x)\,\mathbb{P}(X = x) \\ &= \sum_x \mathbb{P}(X = x)^2.\end{aligned}$$

The variables X and X' take the same value exactly when both equal some value x. Because X and X' are IID, the probability that both equal x is $\mathbb{P}(X = x)\,\mathbb{P}(X' = x) = \mathbb{P}(X = x)^2$. Adding over the values x gives

$$\begin{aligned}\mathbb{P}(X = X') &= \sum_x \mathbb{P}(X = x, X' = x) \\ &= \sum_x \mathbb{P}(X = x)^2.\end{aligned}$$

The expectation $\mathbb{E}[\mathbb{P}(X)]$ and the collision probability $\mathbb{P}(X = X')$ both equal $\sum_x \mathbb{P}(X = x)^2$, so they are equal. □

Theorem 0.6 (The collision entropy is at most the Shannon entropy)**.** *Let X be a discrete random variable with values $x_1, \ldots, x_n$. Then the collision entropy is at most the entropy:*

$$\mathbb{H}_2(X) \leq \mathbb{H}(X).$$

The two coincide exactly when each value that X takes with positive probability has the same probability.

Proof

Proof. Write $\mathbb{P}(X)$ for the probability mass function evaluated at X. The function $-\log$ is strictly convex, so Jensen's inequality gives

$$\mathbb{E}[-\log \mathbb{P}(X)] \geq -\log \mathbb{E}[\mathbb{P}(X)],$$

with equality exactly when $\mathbb{P}(X)$ is constant. The left side is $\mathbb{H}(X)$. The right side is $\mathbb{H}_2(X)$ by its definition. The function $\mathbb{P}(X)$ is constant exactly when every value that X takes with positive probability has that same probability. □

Example 0.8 (Collision entropy and entropy of a three-valued random variable)**.**
Let X take three values with probabilities $1/2$, $1/4$ and $1/4$. The code below computes the expected probability, the collision entropy and the entropy, with base-2 logarithms.

```r
p <- c(1 / 2, 1 / 4, 1 / 4)
expected_prob <- sum(p^2)
collision_entropy <- -log2(expected_prob)
shannon_entropy <- -sum(p * log2(p))
c(
  expected_prob = expected_prob,
  collision_entropy = collision_entropy,
  shannon_entropy = shannon_entropy
)
#>     expected_prob collision_entropy   shannon_entropy
#>          0.375000          1.415037          1.500000
```

By Theorem 0.5, the expected probability $\mathbb{E}[\mathbb{P}(X)] = 0.375$ is the collision probability. The collision entropy is $\mathbb{H}_2(X) = 1.415$ bits. The entropy is $\mathbb{H}(X) = 1.5$ bits. As Theorem 0.6 states, $\mathbb{H}_2(X) \leq \mathbb{H}(X)$, and the two differ because the three values are not equally likely.

Remark 0.11 (Two ways to summarize a random variable by its probabilities). Entropy is the expectation of the surprisal $-\log \mathbb{P}(X)$. Taking the expectation of the probability $\mathbb{P}(X)$ itself summarizes X differently. For a discrete random variable, Theorem 0.5 identifies that expectation with the collision probability, so the collision entropy is $\mathbb{H}_2(X) = -\log \mathbb{P}(X = X')$.
For a continuous random variable with density $\mathbb{p}$, the analogue is $\mathbb{E}[\mathbb{p}(X)] = \int \mathbb{p}(x)^2 \, dx$. The expected-density exercises derive this quantity, then compute it in the uniform case and the normal case.
Theorem 0.6 compares the collision entropy with the Shannon entropy.

Remark 0.12 (Surprisal and the Fisher information). When the probabilities depend on a parameter θ, the surprisal of a value x is $-\log \mathbb{p}(x \mid \theta)$, which is minus the log-likelihood:

$$\mathbb{s}_\theta(x) = -\log \mathbb{p}(x \mid \theta) = -\ell(\theta).$$

Its derivative with respect to θ is minus the score:

$$\frac{\partial}{\partial \theta} \mathbb{s}_\theta(x) = -\ell'(\theta).$$

The expected squared score is the expected information, also called the Fisher information. It is also the expected second derivative of the surprisal, when X is drawn from the model at the true parameter value θ_0, and the score is evaluated at that same X:

$$\mathcal{J}(\theta_0) = \mathbb{E}_{\theta_0}\left[\ell'(\theta_0)^2\right] = \mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\mathfrak{s}_\theta(X)\Big|_{\theta=\theta_0}\right].$$

It is also the second derivative, at the true parameter value, of the Kullback–Leibler divergence between the model at that value and the model at the varying value (proved in the sds notes):

$$\frac{\partial^2}{\partial\theta^2} D_{\mathrm{KL}}(\theta_0 \,\|\, \theta)\Big|_{\theta=\theta_0} = \mathcal{J}(\theta_0).$$

Joint and conditional entropy

Definition 0.8 (Joint entropy)**.** Let X and Y be discrete random variables with joint probability mass function $p(x,y) \overset{\text{def}}{=} \mathbb{P}(X = x, Y = y)$. Their **joint entropy** is the entropy of the pair (X, Y):

$$H(X,Y) \overset{\text{def}}{=} \mathbb{E}[-\log p(X,Y)].$$

Definition 0.9 (Entropy given one value)**.** For X and Y as in Definition 0.8, write $p(y \mid x) \overset{\text{def}}{=} \mathbb{P}(Y = y \mid X = x)$. The **entropy of Y given $X = x$**, for a value x with $\mathbb{P}(X = x) > 0$, is the expectation of $-\log p(Y \mid x)$ given $X = x$:

$$H(Y \mid X = x) \overset{\text{def}}{=} \mathbb{E}[-\log p(Y \mid x) \mid X = x].$$

Definition 0.10 (Conditional entropy)**.** For X and Y as in Definition 0.8 and $p(y \mid x)$ as in Definition 0.9, the **conditional entropy** of Y given X is the expectation of $-\log p(Y \mid X)$:

$$H(Y \mid X) \overset{\text{def}}{=} \mathbb{E}[-\log p(Y \mid X)].$$

Theorem 0.7 (Conditional entropy as an average of entropies given a value)**.** *For X and Y as in Definition 0.8,*

$$H(Y \mid X) = \sum_x p(x)\, H(Y \mid X = x),$$

where $p(x) \overset{\text{def}}{=} \mathbb{P}(X = x)$, and the sum runs over the values x with $p(x) > 0$.

Proof

Proof. The law of iterated expectations conditions on X:

$$\begin{aligned} H(Y \mid X) &= \mathbb{E}[-\log p(Y \mid X)] \\ &= \mathbb{E}[\mathbb{E}[-\log p(Y \mid X) \mid X]] \\ &= \sum_x p(x)\, \mathbb{E}[-\log p(Y \mid x) \mid X = x] \\ &= \sum_x p(x)\, H(Y \mid X = x). \end{aligned}$$

The third line is the definition of expectation over the values x of X, and the last line is Definition 0.9. □

Theorem 0.8 (Chain rule for entropy)**.** *For discrete random variables X and Y,*

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

Proof

Proof. Use the notation of Definition 0.8 and Definition 0.9. With probability 1, $p(X, Y) = p(X)\, p(Y \mid X)$ and $p(X, Y) > 0$, so

$$-\log p(X, Y) = -\log p(X) - \log p(Y \mid X).$$

Both terms on the right are nonnegative, so the expectation of the sum is the sum of the expectation of each term:

$$\begin{aligned} H(X, Y) &= \mathbb{E}[-\log p(X, Y)] \\ &= \mathbb{E}[-\log p(X)] + \mathbb{E}[-\log p(Y \mid X)] \\ &= H(X) + H(Y \mid X). \end{aligned}$$

□

Corollary 0.2 (Chain rule for entropy, other order)**.** *For discrete random variables X and Y,*

$$\mathbb{H}(X, Y) = \mathbb{H}(X \mid Y) + \mathbb{H}(Y).$$

Proof

Proof. The pair (Y, X) takes each value (y, x) with the same probability that (X, Y) takes (x, y), so

$$\mathbb{H}(Y, X) = \mathbb{H}(X, Y).$$

The chain rule, applied with the roles of X and Y exchanged, gives

$$\mathbb{H}(Y, X) = \mathbb{H}(Y) + \mathbb{H}(X \mid Y).$$

Combining the two displays gives the claim. □

Corollary 0.3 (Conditional entropy as a difference)**.** *For discrete random variables X and Y with finite entropies,*

$$\mathbb{H}(X \mid Y) = \mathbb{H}(X, Y) - \mathbb{H}(Y).$$

Proof

Proof. Subtract $\mathbb{H}(Y)$ from both sides of Corollary 0.2:

$$\begin{aligned} \mathbb{H}(X, Y) - \mathbb{H}(Y) &= \mathbb{H}(X \mid Y) + \mathbb{H}(Y) - \mathbb{H}(Y) \\ &= \mathbb{H}(X \mid Y). \end{aligned}$$

□

Example 0.9 (Chain rule for entropy with a noisy copy)**.**
Let X be a fair coin toss, coded 0 or 1. Let Y be a noisy copy of X: it equals X with probability 0.8 and the other value with probability 0.2. The joint distribution gives probability 0.4 to $(0, 0)$ and $(1, 1)$, and 0.1 to $(0, 1)$ and $(1, 0)$. The chunk computes each entropy from its definition, in bits.

```r
entropy <- function(p) -sum(p * log2(p))
joint <- matrix(
  data = c(0.4, 0.1, 0.1, 0.4),
  nrow = 2,
  dimnames = list(x = 0:1, y = 0:1)
)
p_y <- colSums(joint)
cond_x_given_y <- sweep(x = joint, MARGIN = 2, STATS = p_y, FUN = "/")
h_x_given_y_values <- apply(X = cond_x_given_y, MARGIN = 2, FUN = entropy)
h_x_given_y <- sum(p_y * h_x_given_y_values)
h_xy <- entropy(joint)
h_y <- entropy(p_y)
c(h_xy = h_xy, h_x_given_y = h_x_given_y, h_y = h_y)
#>        h_xy h_x_given_y         h_y
#>   1.7219281   0.7219281   1.0000000
```

The chunk computes $\mathbb{H}(X \mid Y)$ as the average of the entropies given each value of Y, as in Theorem 0.7 with X and Y exchanged. Here $\mathbb{H}(X, Y) = 1.722$ bits, $\mathbb{H}(X \mid Y) = 0.722$ bits and $\mathbb{H}(Y) = 1$ bits. So $\mathbb{H}(X \mid Y) + \mathbb{H}(Y) = 1.722$ bits agrees with $\mathbb{H}(X, Y)$ (Corollary 0.2), and $\mathbb{H}(X, Y) - \mathbb{H}(Y) = 0.722$ bits agrees with $\mathbb{H}(X \mid Y)$ (Corollary 0.3).

Definition 0.11 (Joint surprisal)**.** Let X and Y be discrete random variables with joint probability mass function $p(x, y) \stackrel{\text{def}}{=} \mathbb{P}(X = x, Y = y)$. The **joint surprisal** of a pair of values (x, y) with $p(x, y) > 0$ is the surprisal of that pair:

$$\mathbb{s}(x, y) \stackrel{\text{def}}{=} -\log p(x, y).$$

Definition 0.12 (Conditional surprisal)**.** For X and Y as in Definition 0.11, the **conditional surprisal** of a value y given a value x with $\mathbb{P}(X = x) > 0$ and $\mathbb{P}(Y = y \mid X = x) > 0$ is

$$\mathbb{s}(y \mid x) \stackrel{\text{def}}{=} -\log \mathbb{P}(Y = y \mid X = x).$$

Example 0.10 (Joint and conditional surprisal of a noisy copy)**.**
Let X be a fair coin toss, coded 0 or 1. Let Y be a noisy copy of X: it equals X with probability 0.8 and the other value with probability 0.2. Take the pair of values $x = 0$ and $y = 1$. The joint probability is $\mathbb{P}(X = 0, Y = 1) = 0.1$, the probability of $X = 0$ is 0.5, and the conditional probability is $\mathbb{P}(Y = 1 \mid X = 0) = 0.2$. The chunk computes the three surprisals, in bits.

```r
s_xy <- -log2(0.1)
s_x <- -log2(0.5)
s_y_given_x <- -log2(0.2)
c(s_xy = s_xy, s_x = s_x, s_y_given_x = s_y_given_x)
#>        s_xy         s_x s_y_given_x
#>    3.321928    1.000000    2.321928
```

The joint surprisal is $\mathfrak{s}(0,1) = 3.322$ bits, the surprisal of $x = 0$ alone is $\mathfrak{s}(0) = 1$ bit, and the conditional surprisal is $\mathfrak{s}(1 \mid 0) = 2.322$ bits.

Theorem 0.9 (Chain rule for surprisal)**.** *For discrete random variables X and Y and values x and y with $\mathbb{P}(X = x, Y = y) > 0$,*

$$\mathfrak{s}(x, y) = \mathfrak{s}(x) + \mathfrak{s}(y \mid x).$$

Proof

Proof. Since $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y \mid X = x)$,

$$\begin{aligned}
\mathfrak{s}(x, y) &= -\log \mathbb{P}(X = x, Y = y) \\
&= -\log[\mathbb{P}(X = x)\,\mathbb{P}(Y = y \mid X = x)] \\
&= -\log \mathbb{P}(X = x) - \log \mathbb{P}(Y = y \mid X = x) \\
&= \mathfrak{s}(x) + \mathfrak{s}(y \mid x).
\end{aligned}$$

□

Corollary 0.4 (Conditional surprisal as a difference)**.** *For X, Y, x and y as in Theorem 0.9,*

$$\mathfrak{s}(y \mid x) = \mathfrak{s}(x, y) - \mathfrak{s}(x).$$

Proof

Proof. Subtract $\mathfrak{s}(x)$ from both sides of Theorem 0.9:

$$\begin{aligned}
\mathfrak{s}(x, y) - \mathfrak{s}(x) &= \mathfrak{s}(x) + \mathfrak{s}(y \mid x) - \mathfrak{s}(x) \\
&= \mathfrak{s}(y \mid x).
\end{aligned}$$

□

Example 0.11 (Chain rule for surprisal with a noisy copy)**.**
For the pair of values in Example 0.10, the chunk recomputes the three surprisals in bits.

```
s_xy <- -log2(0.1)
s_x <- -log2(0.5)
s_y_given_x <- -log2(0.2)
c(sum = s_x + s_y_given_x, joint = s_xy)
#>      sum    joint
#> 3.321928 3.321928
```

The surprisal of $x = 0$ plus the conditional surprisal of $y = 1$ given $x = 0$ is 3.322 bits. This sum is the joint surprisal $\mathrm{s}(0, 1) = 3.322$ bits (Theorem 0.9). Subtracting $\mathrm{s}(0) = 1$ bit from the joint surprisal recovers $\mathrm{s}(1 \mid 0) = 2.322$ bits (Corollary 0.4).

Lemma 0.1 (Chain rule for surprisal with the conditioning reversed)**.** *For discrete random variables X and Y and outcomes x and y with $\mathbb{P}(X = x, Y = y) > 0$,*

$$\mathrm{s}(x, y) = \mathrm{s}(x \mid y) + \mathrm{s}(y).$$

Proof

Proof. Applying Theorem 0.9 with the roles of X and Y swapped gives

$$\mathrm{s}(y, x) = \mathrm{s}(y) + \mathrm{s}(x \mid y).$$

Both $\mathbb{P}(Y = y, X = x)$ and $\mathbb{P}(X = x, Y = y)$ are the probability of the same pair of values, so $\mathrm{s}(y, x) = \mathrm{s}(x, y)$. Substituting this into the display above, and swapping the two summands, gives the claim. □

Example 0.12 (Chain rule for surprisal with the conditioning reversed, for a noisy copy)**.**

For the pair of values in Example 0.10, $\mathbb{P}(X = 0, Y = 1) = 0.1$, $\mathbb{P}(Y = 1) = 0.5$ and $\mathbb{P}(X = 0 \mid Y = 1) = 0.2$. The chunk computes the three surprisals in bits.

```
s_xy <- -log2(0.1)
s_y <- -log2(0.5)
s_x_given_y <- -log2(0.2)
c(sum = s_x_given_y + s_y, joint = s_xy)
#>      sum    joint
#> 3.321928 3.321928
```

The conditional surprisal of $x = 0$ given $y = 1$ plus the surprisal of $y = 1$ is 3.322 bits. This sum is the joint surprisal $\mathfrak{s}(0, 1) = 3.322$ bits (Lemma 0.1).

Remark 0.13 (The chain rule for entropy is the chain rule for surprisal on average). The entropy of a pair is the expectation of its joint surprisal, and the conditional entropy is the expectation of the conditional surprisal:

$$\mathbb{H}(X, Y) = \mathbb{E}[\mathfrak{s}(X, Y)]$$
$$\mathbb{H}(Y \mid X) = \mathbb{E}[\mathfrak{s}(Y \mid X)].$$

Taking the expectation of both sides of Theorem 0.9 therefore gives Theorem 0.8.

Mutual information

Definition 0.13 (Mutual information)**.** The **mutual information** between discrete random variables X and Y with finite entropies is the sum of their entropies minus their joint entropy:

$$\mathbb{I}(X; Y) \stackrel{\text{def}}{=} \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X, Y).$$

Theorem 0.10 (Mutual information is symmetric)**.** *For discrete random variables X and Y with finite entropies,*

$$\mathbb{I}(X; Y) = \mathbb{I}(Y; X).$$

Proof

Proof. The joint entropy of the pair (X, Y) equals that of (Y, X), because both are the entropy of the same two values, listed in either order. So Definition 0.13 gives

$$\begin{aligned}\mathbb{I}(X; Y) &= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X, Y) \\ &= \mathbb{H}(Y) + \mathbb{H}(X) - \mathbb{H}(Y, X) \\ &= \mathbb{I}(Y; X).\end{aligned}$$

□

Theorem 0.11 (Mutual information as a reduction in entropy)**.** *For discrete random variables X and Y with finite entropies, the mutual information is the reduction in the entropy of either variable from learning the other:*

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

Proof

Proof. Subtract $\mathbb{H}(X)$ from both sides of the chain rule:

$$\mathbb{H}(Y \mid X) = \mathbb{H}(X,Y) - \mathbb{H}(X).$$

Substituting into Definition 0.13 gives

$$\begin{aligned}\mathbb{H}(Y) - \mathbb{H}(Y \mid X) &= \mathbb{H}(Y) - [\mathbb{H}(X,Y) - \mathbb{H}(X)] \\ &= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \\ &= \mathbb{I}(X;Y).\end{aligned}$$

The chain rule also gives $\mathbb{H}(X \mid Y) = \mathbb{H}(X,Y) - \mathbb{H}(Y)$, so exchanging X and Y in the aligned display above gives $\mathbb{H}(X) - \mathbb{H}(X \mid Y) = \mathbb{I}(X;Y)$. □

Theorem 0.12 (Mutual information from the conditional entropies)**.** *For discrete random variables X and Y with finite entropies,*

$$\mathbb{I}(X;Y) = \mathbb{H}(X,Y) - [\mathbb{H}(X \mid Y) + \mathbb{H}(Y \mid X)].$$

Proof

Proof. By Corollary 0.3 and the chain rule, the entropy of each variable is the joint entropy minus a conditional entropy:

$$\begin{aligned}\mathbb{H}(X) &= \mathbb{H}(X,Y) - \mathbb{H}(Y \mid X) \\ \mathbb{H}(Y) &= \mathbb{H}(X,Y) - \mathbb{H}(X \mid Y).\end{aligned}$$

Substituting both into Definition 0.13 gives

$$\begin{aligned}\mathbb{I}(X;Y) &= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \\ &= [\mathbb{H}(X,Y) - \mathbb{H}(Y \mid X)] + [\mathbb{H}(X,Y) - \mathbb{H}(X \mid Y)] - \mathbb{H}(X,Y) \\ &= \mathbb{H}(X,Y) - [\mathbb{H}(X \mid Y) + \mathbb{H}(Y \mid X)].\end{aligned}$$

□

Theorem 0.13 (Decomposition of the sum of entropies)**.** *For discrete random variables X and Y with finite entropies,*

$$\mathbb{H}(X) + \mathbb{H}(Y) = \mathbb{H}(X \mid Y) + 2\,\mathbb{I}(X;Y) + \mathbb{H}(Y \mid X).$$

Proof

Proof. Rearranging Theorem 0.12 gives

$$\mathbb{H}(X,Y) = \mathbb{I}(X;Y) + \mathbb{H}(X \mid Y) + \mathbb{H}(Y \mid X).$$

Rearranging Definition 0.13 gives

$$\mathbb{H}(X) + \mathbb{H}(Y) = \mathbb{I}(X;Y) + \mathbb{H}(X,Y).$$

Substituting the first display into the second gives

$$\begin{aligned}\mathbb{H}(X) + \mathbb{H}(Y) &= \mathbb{I}(X;Y) + [\mathbb{I}(X;Y) + \mathbb{H}(X \mid Y) + \mathbb{H}(Y \mid X)] \\ &= \mathbb{H}(X \mid Y) + 2\,\mathbb{I}(X;Y) + \mathbb{H}(Y \mid X).\end{aligned}$$

□

Theorem 0.14 (Twice the mutual information)**.** *For discrete random variables X and Y with finite entropies,*

$$2\,\mathbb{I}(X;Y) = [\mathbb{H}(X) - \mathbb{H}(X \mid Y)] + [\mathbb{H}(Y) - \mathbb{H}(Y \mid X)].$$

Proof

Proof. By Theorem 0.11,

$$\begin{aligned}\mathbb{H}(X) - \mathbb{H}(X \mid Y) &= \mathbb{I}(X;Y) \\ \mathbb{H}(Y) - \mathbb{H}(Y \mid X) &= \mathbb{I}(X;Y).\end{aligned}$$

Adding the two displays gives

$$\begin{aligned}[\mathbb{H}(X) - \mathbb{H}(X \mid Y)] + [\mathbb{H}(Y) - \mathbb{H}(Y \mid X)] &= \mathbb{I}(X;Y) + \mathbb{I}(X;Y) \\ &= 2\,\mathbb{I}(X;Y).\end{aligned}$$

□

Example 0.13 (Mutual information identities with a noisy copy)**.**
Take the fair coin X and its noisy copy Y from Example 0.9. The chunk computes the entropies from the joint probabilities in bits, then the mutual information in four ways.

```r
entropy <- function(p) -sum(p * log2(p))
joint <- matrix(
  data = c(0.4, 0.1, 0.1, 0.4),
  nrow = 2
)
h_xy <- entropy(joint)
h_x <- entropy(rowSums(joint))
h_y <- entropy(colSums(joint))
h_x_given_y <- h_xy - h_y
h_y_given_x <- h_xy - h_x
mi_definition <- h_x + h_y - h_xy
mi_reduction <- h_y - h_y_given_x
mi_conditionals <- h_xy - (h_x_given_y + h_y_given_x)
mi_twice <- ((h_x - h_x_given_y) + (h_y - h_y_given_x)) / 2
c(mi_definition, mi_reduction, mi_conditionals, mi_twice)
#> [1] 0.2780719 0.2780719 0.2780719 0.2780719
```

All four agree: $\mathbb{I}(X;Y) = 0.278$ bits (Definition 0.13, Theorem 0.11, Theorem 0.12 and Theorem 0.14). Also $\mathbb{H}(X)+\mathbb{H}(Y) = 2$ bits matches $\mathbb{H}(X \mid Y)+2\,\mathbb{I}(X;Y)+\mathbb{H}(Y \mid X) = 2$ bits (Theorem 0.13).

Remark 0.14 (Entropies as areas). The identities above match a picture of two overlapping regions, with the area of one region standing for $\mathbb{H}(X)$ and the area of the other for $\mathbb{H}(Y)$. The overlap has area $\mathbb{I}(X;Y)$ (Definition 0.13). The part of the first region outside the overlap has area $\mathbb{H}(X \mid Y)$, and the part of the second region outside the overlap has area $\mathbb{H}(Y \mid X)$. The two regions together have area $\mathbb{H}(X,Y)$ (Theorem 0.12). The picture is a mnemonic for two variables only.

Theorem 0.15 (Mutual information as a divergence)**.** *Let X and Y be discrete random variables with joint probability mass function $p(x,y)$ and marginal probability mass functions $p(x)$ and $p(y)$. If their entropies are finite, then their mutual information is the Kullback–Leibler divergence of the product of the marginals from the joint distribution:*

$$\mathbb{I}(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)},$$

where the sum runs over the pairs with $p(x,y) > 0$.

Proof

Proof. The entropies are finite, so the expectations below are finite and can be split and regrouped. With probability 1, $p(X, Y) > 0$, $p(X) > 0$ and $p(Y) > 0$. By Definition 0.13 and the definitions of entropy and joint entropy,

$$\begin{aligned}
\mathbb{I}(X;Y) &= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \\
&= \mathbb{E}[-\log p(X)] + \mathbb{E}[-\log p(Y)] - \mathbb{E}[-\log p(X,Y)] \\
&= \mathbb{E}[\log p(X,Y) - \log p(X) - \log p(Y)] \\
&= \mathbb{E}\left[\log \frac{p(X,Y)}{p(X)\,p(Y)}\right] \\
&= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}.
\end{aligned}$$

The last line is the definition of expectation over the pairs with $p(x, y) > 0$. It is the divergence of $p(x)\,p(y)$ from $p(x, y)$ by the discrete formula. □

Example 0.14 (Mutual information of a noisy copy)**.**
Let X be a fair coin toss, coded 0 or 1. Let Y equal X with probability 0.8 and the other value with probability 0.2. The joint distribution gives probability 0.4 to $(0, 0)$ and $(1, 1)$ and 0.1 to $(0, 1)$ and $(1, 0)$. The chunk computes the mutual information in bits from the divergence formula in Theorem 0.15.

```
joint <- matrix(
  data = c(0.4, 0.1, 0.1, 0.4),
  nrow = 2,
  dimnames = list(x = 0:1, y = 0:1)
)
p_x <- rowSums(joint)
p_y <- colSums(joint)
product <- outer(X = p_x, Y = p_y)
terms <- joint * log2(joint / product)
mutual_information <- sum(terms)
mutual_information
#> [1] 0.2780719
```

So $\mathbb{I}(X;Y) = 0.278$ bits. This matches $\mathbb{H}(Y) - \mathbb{H}(Y \mid X)$, which is 1 bit minus the entropy of a coin with probabilities 0.8 and 0.2. That is $1 - 0.722$ bits.

Theorem 0.16 (The divergence is nonnegative)**.** *Let p and q be probability mass functions on the values $x_1, x_2, \ldots$, with $q(x) > 0$ wherever $p(x) > 0$, and suppose that*

$$\sum_x p(x) \left|\log_2(p(x)/q(x))\right| < \infty.$$

Then the Kullback–Leibler divergence, with base-2 logarithms, satisfies

$$D_{\mathrm{KL}}(p \,\|\, q) \geq 0,$$

with $D_{\mathrm{KL}}(p \,\|\, q) = 0$ exactly when $q = p$.

Proof

Proof. Every sum here runs over the values x with $p(x) > 0$. For every $t > 0$, the logarithm lies below its tangent line: $\ln t \leq t - 1$, with equality exactly when $t = 1$. Apply this with $t = q(x)/p(x)$, and multiply by $p(x) > 0$, as in the proof of Gibbs' inequality for finitely many values:

$$p(x) \ln \frac{q(x)}{p(x)} \leq q(x) - p(x),$$

with equality exactly when $q(x) = p(x)$. Both sides are summable over x. The left side does by the assumption in the statement, and the right side does because $\sum_x p(x) = 1$ and $\sum_x q(x) \leq 1$. So the relation holds for the totals over x:

$$\sum_x p(x) \ln \frac{q(x)}{p(x)} \leq \sum_x q(x) - 1 \leq 0.$$

The left side is $-\ln 2 \cdot D_{\mathrm{KL}}(p \,\|\, q)$, because the divergence uses base-2 logarithms, so $D_{\mathrm{KL}}(p \,\|\, q) \geq 0$.
If $D_{\mathrm{KL}}(p \,\|\, q) = 0$, both relations hold with equality. The first forces $q(x) = p(x)$ at every x with $p(x) > 0$. The second then gives $\sum_x q(x) = 1$, so q has no mass where p has none, and $q = p$. Conversely, $q = p$ gives $\log_2(p(x)/q(x)) = 0$ for every x. □

Example 0.15 (Divergence of a biased coin from a fair coin)**.**
Let $p = (0.9, 0.1)$ be the probability mass function of a biased coin, and let $q = (0.5, 0.5)$ be a fair coin. The code below computes the divergence both ways, in bits.

```r
p <- c(0.9, 0.1)
q <- c(0.5, 0.5)
c(
  p_from_q = sum(p * log2(p / q)),
  q_from_p = sum(q * log2(q / p)),
  p_from_p = sum(p * log2(p / p))
)
#>  p_from_q  q_from_p  p_from_p
#> 0.5310044 0.7369656 0.0000000
```

Both divergences between the two coins are positive, as Theorem 0.16 requires, and they differ, so swapping p and q changes the divergence. The divergence of p from itself is 0.

Theorem 0.17 (Mutual information is nonnegative)**.** *For discrete random variables X and Y with finite entropies,*

$$\mathbb{I}(X;Y) \geq 0.$$

Equality holds exactly when X and Y are independent.

Proof

Proof. Let $p(x, y)$ be the joint mass function, with marginals $p(x)$ and $p(y)$. By Theorem 0.15, $\mathbb{I}(X;Y)$ is the divergence of $p(x)\,p(y)$ from $p(x, y)$, and the sum defining it is absolutely summable, as the proof of that result shows. The product $p(x)\,p(y)$ is nonzero wherever $p(x, y)$ is nonzero, because each of $p(x)$ and $p(y)$ is at least $p(x, y)$ there. By Theorem 0.16, the divergence is nonnegative, with equality exactly when $p(x)\,p(y)$ and $p(x, y)$ agree. The logarithm satisfies $\log_b t = \ln t / \ln b$, so the divergence with $\log_b$ is the divergence with $\log_2$ times $\ln 2 / \ln b$, which has the same sign and the same zero. Hence $\mathbb{I}(X;Y) \geq 0$, and $\mathbb{I}(X;Y) = 0$ exactly when $p(x, y) = p(x)\,p(y)$ for every pair, which is independence. □

Example 0.16 (Mutual information of dependent and independent coins)**.**
For the noisy copy in Example 0.14, the mutual information is positive, because Y depends on X. For two independent fair coins, every joint probability is 0.25. The chunk computes the mutual information of both pairs.

```r
joint_dependent <- matrix(data = c(0.4, 0.1, 0.1, 0.4), nrow = 2)
ratio_dependent <- joint_dependent / outer(
  X = rowSums(joint_dependent),
  Y = colSums(joint_dependent)
)
mi_dependent <- sum(joint_dependent * log2(ratio_dependent))
joint_independent <- matrix(data = 0.25, nrow = 2, ncol = 2)
product_independent <- outer(
  X = rowSums(joint_independent),
  Y = colSums(joint_independent)
)
ratio_independent <- joint_independent / product_independent
terms_independent <- joint_independent * log2(ratio_independent)
mi_independent <- sum(terms_independent)
c(mi_dependent = mi_dependent, mi_independent = mi_independent)
#>   mi_dependent mi_independent
#>      0.2780719      0.0000000
```

So the mutual information of the independent coins is 0 bits, and that of the dependent pair is 0.278 bits (Theorem 0.17).

Corollary 0.5 (Conditioning reduces entropy)**.** *For discrete random variables X and Y with finite entropies,*

$$\mathbb{H}(Y \mid X) \leq \mathbb{H}(Y).$$

Equality holds exactly when X and Y are independent.

Proof

Proof. By Theorem 0.11,

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

By Theorem 0.17, the right side is nonnegative, and it is 0 exactly when X and Y are independent. □

Example 0.17 (Learning a noisy copy lowers the entropy)**.**
For the coin X and noisy copy Y in Example 0.14, the entropy of Y is 1 bit. Given X, the variable Y is a coin with probabilities 0.8 and 0.2. The chunk computes the conditional entropy from that coin.

```r
h_y <- 1
h_y_given_x <- -0.8 * log2(0.8) - 0.2 * log2(0.2)
c(h_y = h_y, h_y_given_x = h_y_given_x)
#>         h_y h_y_given_x
#>   1.0000000   0.7219281
```

Here $\mathbb{H}(Y \mid X) = 0.722$ bits, which is below $\mathbb{H}(Y) = 1$ bit (Corollary 0.5).

Remark 0.15 (Mutual information as a divergence). Cover and Thomas (2006, sec. 2.3) define mutual information as the Kullback–Leibler divergence of the product of the marginals of (X, Y) from their joint distribution. Theorem 0.15 proves that this divergence equals $\mathbb{H}(Y) - \mathbb{H}(Y \mid X)$ (Cover and Thomas 2006, sec. 2.4), and Theorem 0.17 proves that $\mathbb{I}(X; Y) \geq 0$ (Cover and Thomas 2006, sec. 2.6). Learning X never raises the entropy of Y on average (Corollary 0.5).

Theorem 0.18 (Mutual information of a variable with itself)**.** *For a discrete random variable X with finite entropy,*

$$\mathbb{I}(X; X) = \mathbb{H}(X).$$

Proof

Proof. Given $X = x$, the variable X equals x with probability 1, so $p(x \mid x) = 1$. Hence $-\log p(X \mid x) = -\log 1 = 0$ given $X = x$, and the expectation of this constant is 0:

$$\mathbb{H}(X \mid X = x) = 0.$$

By Theorem 0.7,

$$\begin{aligned} \mathbb{H}(X \mid X) &= \sum_x p(x)\, \mathbb{H}(X \mid X = x) \\ &= 0. \end{aligned}$$

By Theorem 0.11 with $Y = X$,

$$\begin{aligned} \mathbb{I}(X; X) &= \mathbb{H}(X) - \mathbb{H}(X \mid X) \\ &= \mathbb{H}(X). \end{aligned}$$

□

Example 0.18 (A fair coin is fully informative about itself)**.**
Let X be a fair coin toss. Its entropy is 1 bit. Given X itself, there is no uncertainty left, so $\mathbb{H}(X \mid X) = 0$. Hence $\mathbb{I}(X;X) = 1 - 0 = 1$ bit (Theorem 0.18), the same as $\mathbb{H}(X)$.

Definition 0.14 (Information gain)**.** A decision tree splits the points at a node into branches. The **information gain** of the split is the entropy of the labels at the node, minus the entropy of the labels in each branch, weighted by the share of the node's points that branch holds (Quinlan 1986).

Remark 0.16 (Information gain is mutual information)*.* Draw one point from the node of Definition 0.14, and let Y be its class and X the branch it falls into. Then the entropy at the node is $H(Y)$, the weighted entropy left in the branches is $H(Y \mid X)$, and the information gain is the mutual information $\mathbb{I}(X;Y)$.

Exercise 0.1 (Choosing between two splits)**.** A node holds 20 points, 10 red and 10 green. Two candidate splits each send the points into two branches:

split	branch	red	green
A	yes	4	0
	no	6	10
B	yes	5	10
	no	5	0

Draw one point from the node, each equally likely, let Y be its color, and let X be the branch it falls into (Remark 0.16). The information gain of each split is then $\mathbb{I}(X;Y)$. Use base-2 logarithms throughout.

1. Compute $H(Y)$.
2. For each split, compute $H(Y \mid X = x)$ for both branches, and then $H(Y \mid X)$.
3. Compute the information gain $\mathbb{I}(X;Y)$ for each split. Which split has the larger gain?

Solution

Solution 0.1*.*

1. Half the points are red, so

$$\begin{aligned} H(Y) &= -\tfrac{1}{2}\log_2 \tfrac{1}{2} - \tfrac{1}{2}\log_2 \tfrac{1}{2} \\ &= 1 \text{ bit.} \end{aligned}$$

2. A branch holding one color only has entropy 0. For split A, the "no" branch holds 6 red and 10 green points of 16:

$$\begin{aligned} H(Y \mid X = \text{no}) &= -\tfrac{6}{16}\log_2 \tfrac{6}{16} - \tfrac{10}{16}\log_2 \tfrac{10}{16} \\ &\approx 0.531 + 0.424 \\ &\approx 0.954 \text{ bits.} \end{aligned}$$

   The "yes" branch holds 4 of the 20 points and the "no" branch 16, so

$$\begin{aligned} H(Y \mid X) &= \tfrac{4}{20} \cdot 0 + \tfrac{16}{20} \cdot 0.954 \\ &\approx 0.764 \text{ bits.} \end{aligned}$$

   For split B, the "yes" branch holds 5 red and 10 green points of 15:

$$\begin{aligned} H(Y \mid X = \text{yes}) &= -\tfrac{1}{3}\log_2 \tfrac{1}{3} - \tfrac{2}{3}\log_2 \tfrac{2}{3} \\ &\approx 0.528 + 0.390 \\ &\approx 0.918 \text{ bits.} \end{aligned}$$

   The "yes" branch holds 15 of the 20 points and the "no" branch 5, so

$$\begin{aligned} H(Y \mid X) &= \tfrac{15}{20} \cdot 0.918 + \tfrac{5}{20} \cdot 0 \\ &\approx 0.689 \text{ bits.} \end{aligned}$$

3. By Theorem 0.11, the gain is $H(Y) - H(Y \mid X)$:

   - split A: $1 - 0.764 = 0.236$ bits
   - split B: $1 - 0.689 = 0.311$ bits

   Split B has the larger gain. Both splits isolate a branch of one color, so the gain depends on the other, mixed branch. B's mixed branch, 5 red against 10 green, is more lopsided than A's 6 against 10, so it has less entropy, and it also holds fewer of the points, 15 against 16, so its entropy counts for less.

Chain rules

Theorem 0.19 (Chain rule for entropy, many variables)**.** *For discrete random variables $X_1, \dots, X_n$ with finite entropies,*

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

where the term for $i = 1$ is $\mathbb{H}(X_1)$.

Proof

Proof. The proof is by induction on n. The claim for $n = 1$ is $\mathbb{H}(X_1) = \mathbb{H}(X_1)$.
Let $n \geq 2$, and suppose the claim holds for $n - 1$ variables. Treat $(X_1, \dots, X_{n-1})$ as a single variable. The chain rule for two variables (Theorem 0.8) gives

$$\mathbb{H}(X_1, \dots, X_n) = \mathbb{H}(X_1, \dots, X_{n-1}) + \mathbb{H}(X_n \mid X_1, \dots, X_{n-1}).$$

The induction hypothesis expands the first summand on the right:

$$\begin{aligned} \mathbb{H}(X_1, \dots, X_n) &= \sum_{i=1}^{n-1} \mathbb{H}(X_i \mid X_1, \dots, X_{i-1}) + \mathbb{H}(X_n \mid X_1, \dots, X_{n-1}) \\ &= \sum_{i=1}^{n} \mathbb{H}(X_i \mid X_1, \dots, X_{i-1}). \end{aligned}$$

□

Example 0.19 (Chain rule for entropy along a chain of noisy copies)**.**
Let X_1 be a fair coin toss, coded 0 or 1. Let X_2 be a noisy copy of X_1 that flips it with probability 0.2, and let X_3 be a noisy copy of X_2 that flips it with probability 0.2, independently of X_1 given X_2. The chunk builds the joint probability mass function of (X_1, X_2, X_3) and compares their entropy, in bits, with the sum in Theorem 0.19.

```r
flip <- 0.2
channel <- matrix(
  data = c(1 - flip, flip, flip, 1 - flip),
  nrow = 2,
  byrow = TRUE
)
joint <- array(data = 0, dim = c(2, 2, 2))
for (x1 in 1:2) {
  for (x2 in 1:2) {
    for (x3 in 1:2) {
      joint[x1, x2, x3] <- 0.5 * channel[x1, x2] * channel[x2, x3]
    }
  }
}
h_joint <- -sum(joint * log2(joint))
h_coin <- -flip * log2(flip) - (1 - flip) * log2(1 - flip)
h_sum <- 1 + h_coin + h_coin
c(h_joint = h_joint, h_sum = h_sum)
#>  h_joint    h_sum
#> 2.443856 2.443856
```

The first term of the sum is $\mathbb{H}(X_1) = 1$ bit. Given X_1, the variable X_2 is a coin with flip probability 0.2, and given (X_1, X_2) the variable X_3 is the same kind of coin, so each of the other two terms is 0.722 bits. The sum is 2.444 bits, which equals $\mathbb{H}(X_1, X_2, X_3) = 2.444$ bits.

Definition 0.15 (Conditional mutual information)**.** The **conditional mutual information** of discrete random variables X and Y given a third one, Z, with finite entropies, is the reduction in the conditional entropy of Y given Z from also learning X:

$$\mathbb{I}(X; Y \mid Z) \stackrel{\text{def}}{=} \mathbb{H}(Y \mid Z) - \mathbb{H}(Y \mid X, Z).$$

Example 0.20 (Conditional mutual information along a chain of noisy copies)**.**
Take the chain of noisy copies X_1, X_2, X_3 from Example 0.19. Once X_2 is known, X_1 adds nothing to the prediction of X_3, so $\mathbb{H}(X_3 \mid X_1, X_2) = \mathbb{H}(X_3 \mid X_2)$, and both equal the entropy h of a coin with flip probability 0.2. The chunk computes the conditional mutual information.

```
flip <- 0.2
h_coin <- -flip * log2(flip) - (1 - flip) * log2(1 - flip)
h_x3_given_x2 <- h_coin
h_x3_given_x1_x2 <- h_coin
h_x3_given_x2 - h_x3_given_x1_x2
#> [1] 0
```

So $\mathbb{I}(X_1; X_3 \mid X_2) = 0$ bits.

Theorem 0.20 (Chain rule for mutual information)**.** *For discrete random variables X, Y and Z whose entropies $\mathbb{H}(X)$, $\mathbb{H}(Y)$ and $\mathbb{H}(Z)$ are finite,*

$$\mathbb{I}(X, Z; Y) = \mathbb{I}(Z; Y) + \mathbb{I}(X; Y \mid Z).$$

Proof

Proof. By Theorem 0.11, applied to the pair (X, Z) and to Z,

$$\begin{aligned} \mathbb{I}(X, Z; Y) &= \mathbb{H}(Y) - \mathbb{H}(Y \mid X, Z) \\ \mathbb{I}(Z; Y) &= \mathbb{H}(Y) - \mathbb{H}(Y \mid Z). \end{aligned}$$

Subtracting the second display from the first gives

$$\begin{aligned} \mathbb{I}(X, Z; Y) - \mathbb{I}(Z; Y) &= \mathbb{H}(Y \mid Z) - \mathbb{H}(Y \mid X, Z) \\ &= \mathbb{I}(X; Y \mid Z), \end{aligned}$$

where the last line is Definition 0.15. □

Example 0.21 (Chain rule for mutual information along a chain of noisy copies)**.** For the chain in Example 0.19, the marginal distribution of X_3 is a fair coin, so $\mathbb{H}(X_3) = 1$ bit. The chunk computes the mutual information of X_3 with X_2, and with the pair (X_1, X_2).

```
flip <- 0.2
h_coin <- -flip * log2(flip) - (1 - flip) * log2(1 - flip)
h_x3_given_x1_x2 <- h_coin
i_x2_x3 <- 1 - h_coin
i_pair_x3 <- 1 - h_x3_given_x1_x2
c(i_x2_x3 = i_x2_x3, i_pair_x3 = i_pair_x3)
#>   i_x2_x3 i_pair_x3
#> 0.2780719 0.2780719
```

Here $\mathbb{I}(X_2; X_3) = 0.278$ bits and $\mathbb{I}(X_1, X_2; X_3) = 0.278$ bits. The difference is $\mathbb{I}(X_1; X_3 \mid X_2) = 0$ (Theorem 0.20), as in Example 0.20.

Theorem 0.21 (Conditional mutual information is nonnegative)**.** *For discrete random variables X, Y and Z with finite entropies,*

$$\mathbb{I}(X; Y \mid Z) \geq 0.$$

Proof

Proof. For each value z with $\mathbb{P}(Z = z) > 0$, let $\mathbb{I}_z(X; Y)$ be the mutual information of X and Y computed from their joint probabilities given $Z = z$, so that $\mathbb{I}_z(X; Y) = \mathbb{H}(Y \mid Z = z) - \mathbb{H}(Y \mid X, Z = z)$. Applying Definition 0.10 to $\mathbb{H}(Y \mid Z)$ and to $\mathbb{H}(Y \mid X, Z)$ expands each over the values z of Z:

$$\begin{aligned} \mathbb{I}(X; Y \mid Z) &= \mathbb{H}(Y \mid Z) - \mathbb{H}(Y \mid X, Z) \\ &= \sum_z \mathbb{P}(Z = z)[\mathbb{H}(Y \mid Z = z) - \mathbb{H}(Y \mid X, Z = z)] \\ &= \sum_z \mathbb{P}(Z = z)\, \mathbb{I}_z(X; Y). \end{aligned}$$

Each $\mathbb{I}_z(X; Y)$ is nonnegative by Theorem 0.17. Since $\mathbb{P}(Z = z) \geq 0$ as well, every summand $\mathbb{P}(Z = z)\, \mathbb{I}_z(X; Y)$ is nonnegative, so $\mathbb{I}(X; Y \mid Z) \geq 0$. □

Example 0.22 (Learning a third variable can create information)**.**
Let X and Y be independent fair coins, coded 0 or 1, and let $Z = X + Y \bmod 2$ record whether they differ. The two coins are independent, so $\mathbb{I}(X; Y) = 0$ (Theorem 0.17). Given Z, each of X and Y determines the other. The chunk computes the conditional mutual information from the joint probability mass function, in bits.

```r
entropy <- function(p) -sum(p[p > 0] * log2(p[p > 0]))
joint <- array(data = 0, dim = c(2, 2, 2))
for (x in 0:1) {
  for (y in 0:1) {
    z <- (x + y) %% 2
    joint[x + 1, y + 1, z + 1] <- 0.25
  }
}
h_xyz <- entropy(joint)
h_xz <- entropy(apply(X = joint, MARGIN = c(1, 3), FUN = sum))
h_yz <- entropy(apply(X = joint, MARGIN = c(2, 3), FUN = sum))
h_z <- entropy(apply(X = joint, MARGIN = 3, FUN = sum))
h_y_given_z <- h_yz - h_z
h_y_given_xz <- h_xyz - h_xz
h_y_given_z - h_y_given_xz
#> [1] 1
```

So $\mathbb{I}(X;Y \mid Z) = 1$ bit, which is nonnegative (Theorem 0.21), although $\mathbb{I}(X;Y) = 0$.

Definition 0.16 (Markov chain of three random variables)**.** Discrete random variables X, Y and Z form a **Markov chain** $X \to Y \to Z$ (Cover and Thomas 2006, sec. 2.8) if X and Z are conditionally independent given Y: for every value y with $\mathbb{P}(Y = y) > 0$ and all values x and z,

$$\mathbb{P}(X = x, Z = z \mid Y = y) = \mathbb{P}(X = x \mid Y = y)\, \mathbb{P}(Z = z \mid Y = y).$$

Example 0.23 (A chain of noisy copies is a Markov chain)**.**
In Example 0.19, each of X_2 and X_3 is a noisy copy of the variable before it. So X_3 depends on X_1 only through X_2. The chunk builds the joint probability mass function of (X_1, X_2, X_3) and checks the defining equation of the chain $X_1 \to X_2 \to X_3$ for each value of X_2.

```r
flip <- 0.2
copy <- matrix(
  data = c(1 - flip, flip, flip, 1 - flip),
  nrow = 2
)
joint <- array(data = 0, dim = c(2, 2, 2))
for (x1 in 1:2) {
  for (x2 in 1:2) {
    for (x3 in 1:2) {
      joint[x1, x2, x3] <- 0.5 * copy[x1, x2] * copy[x2, x3]
    }
  }
}
gaps <- numeric(2)
for (x2 in 1:2) {
  slice <- joint[, x2, ]
  given_x2 <- slice / sum(slice)
  product <- outer(rowSums(given_x2), colSums(given_x2))
  gaps[x2] <- max(abs(given_x2 - product))
}
gaps
#> [1] 1.110223e-16 1.110223e-16
```

Both gaps are zero up to rounding, so X_1 and X_3 are conditionally independent given X_2.

Lemma 0.2 (Conditional mutual information along a Markov chain)**.** *If $X \to Y \to Z$ is a Markov chain and the entropies of X, Y and Z are finite, then*

$$\mathbb{I}(Z; X \mid Y) = 0.$$

Proof

Proof. Fix values y and z with $\mathbb{P}(Y = y, Z = z) > 0$. Dividing the defining equation of the chain by $\mathbb{P}(Z = z \mid Y = y)$ gives

$$\mathbb{P}(X = x \mid Y = y, Z = z) = \mathbb{P}(X = x \mid Y = y)$$

for every x. So the entropy given $(Y, Z) = (y, z)$ is the entropy given $Y = y$. Averaging over (Y, Z) and over Y gives

$$\mathbb{H}(X \mid Y, Z) = \mathbb{H}(X \mid Y).$$

By Definition 0.15,

$$\begin{aligned}\mathbb{I}(Z; X \mid Y) &= \mathbb{H}(X \mid Y) - \mathbb{H}(X \mid Z, Y) \\ &= \mathbb{H}(X \mid Y) - \mathbb{H}(X \mid Y, Z) \\ &= 0.\end{aligned}$$

□

Theorem 0.22 (Data processing inequality)**.** *If $X \to Y \to Z$ is a Markov chain and the entropies of X, Y and Z are finite, then*

$$\mathbb{I}(X; Z) \leq \mathbb{I}(X; Y).$$

Proof

Proof. The chain rule for the pair (Z, Y) and the pair (Y, Z) gives

$$\begin{aligned}\mathbb{I}(Z, Y; X) &= \mathbb{I}(Y; X) + \mathbb{I}(Z; X \mid Y) \\ \mathbb{I}(Y, Z; X) &= \mathbb{I}(Z; X) + \mathbb{I}(Y; X \mid Z).\end{aligned}$$

The pairs (Z, Y) and (Y, Z) contain the same two variables, so $\mathbb{I}(Z, Y; X) = \mathbb{I}(Y, Z; X)$. By Lemma 0.2, $\mathbb{I}(Z; X \mid Y) = 0$. By Theorem 0.21, $\mathbb{I}(Y; X \mid Z) \geq 0$. Therefore

$$\begin{aligned}\mathbb{I}(Y; X) &= \mathbb{I}(Z; X) + \mathbb{I}(Y; X \mid Z) \\ &\geq \mathbb{I}(Z; X).\end{aligned}$$

By Theorem 0.10, $\mathbb{I}(Y; X) = \mathbb{I}(X; Y)$ and $\mathbb{I}(Z; X) = \mathbb{I}(X; Z)$, so $\mathbb{I}(X; Z) \leq \mathbb{I}(X; Y)$. □

Example 0.24 (Each extra noisy copy loses information)**.**
For the chain $X_1 \to X_2 \to X_3$ in Example 0.23, the chunk computes the mutual information of X_1 with X_2 and with X_3, in bits. Each variable is a fair coin. The variable X_3 differs from X_1 when exactly one of the two copies flips, which has probability $2 \times 0.2 \times 0.8 = 0.32$.

```
h_coin <- function(flip) -flip * log2(flip) - (1 - flip) * log2(1 - flip)
i_x1_x2 <- 1 - h_coin(flip = 0.2)
i_x1_x3 <- 1 - h_coin(flip = 0.32)
c(i_x1_x2 = i_x1_x2, i_x1_x3 = i_x1_x3)
#>    i_x1_x2    i_x1_x3
#> 0.27807191 0.09561854
```

Here $\mathbb{I}(X_1; X_3) = 0.096$ bits is below $\mathbb{I}(X_1; X_2) = 0.278$ bits (Theorem 0.22).

Stationary sequences

Definition 0.17 (Stationary sequence)**.** A **stationary sequence** is a sequence $X_1, X_2, \ldots$ of discrete random variables whose joint probability mass functions do not change when the sequence is shifted in time: for every length $k \geq 1$, every shift $s \geq 0$ and all values $x_1, \ldots, x_k$,

$$\mathbb{P}(X_{1+s} = x_1, \ldots, X_{k+s} = x_k) = \mathbb{P}(X_1 = x_1, \ldots, X_k = x_k).$$

Example 0.25 (A sticky coin)**.**
Let X_1 be a fair coin toss, coded 0 or 1. For each $t \geq 2$, let X_t equal X_{t-1} with probability 0.9 and $1 - X_{t-1}$ with probability 0.1, independently of $X_1, \ldots, X_{t-2}$ given X_{t-1}. If X_{t-1} is fair, then so is X_t, because

$$\mathbb{P}(X_t = 1) = 0.9 \cdot \tfrac{1}{2} + 0.1 \cdot \tfrac{1}{2},$$

which is $1/2$. So, by induction on t, every X_t is 0 or 1 with probability $1/2$. The probability that k consecutive tosses take given values is then $1/2$ times 0.9 for each repeat and 0.1 for each switch. That product does not depend on which toss the run starts at, so the sticky coin is a stationary sequence.

Example 0.26 (An alternating sequence that starts at zero)**.**
Let $X_1 = 0$, and let $X_t = 1 - X_{t-1}$ for each $t \geq 2$, so the sequence is $0, 1, 0, 1, \ldots$ with probability 1. The sequence is not a stationary sequence: with length $k = 1$ and shift $s = 1$,

$$\mathbb{P}(X_2 = 0) = 0,$$

but

$$\mathbb{P}(X_1 = 0) = 1.$$

Definition 0.18 (N-gram model)**.** An **N-gram model**, for an integer $N \geq 1$, is a sequence $X_1, X_2, \ldots$ of discrete random variables in which each variable depends on the earlier ones only through the $N-1$ variables just before it: for every $t > N$ and all values $x_1, \ldots, x_t$ with $\mathbb{P}(X_1 = x_1, \ldots, X_{t-1} = x_{t-1}) > 0$,

$$\mathbb{P}(X_t = x_t \mid X_1 = x_1, \dots, X_{t-1} = x_{t-1}) = \mathbb{P}(X_t = x_t \mid X_{t-N+1} = x_{t-N+1}, \dots, X_{t-1} = x_{t-1}).$$

Example 0.27 (The sticky coin is a 2-gram model)**.**
In the sticky coin, given all the earlier tosses, X_t repeats X_{t-1} with probability 0.9. That conditional probability involves only the one toss just before X_t, so the sticky coin is a 2-gram model (Definition 0.18).

Example 0.28 (A sequence that repeats every second value)**.**
Let X_1 and X_2 be independent fair coin tosses, coded 0 or 1, and let $X_t = X_{t-2}$ for each $t \geq 3$. The sequence is not a 2-gram model (Definition 0.18). Given both earlier tosses, X_3 is fixed:

$$\mathbb{P}(X_3 = 0 \mid X_1 = 0, X_2 = 0) = 1.$$

Given X_2 alone, $X_3 = X_1$ is still a fair coin, because X_1 is independent of X_2:

$$\mathbb{P}(X_3 = 0 \mid X_2 = 0) = \frac{1}{2}.$$

The sequence is a 3-gram model, because X_t is fixed by X_{t-2}, one of the two variables just before it.

Remark 0.17 (Guessing a text backward). A reader can guess a text one character at a time from its start, guessing each character from the characters already read. The reader can also guess it from its end, guessing each character from the characters that follow it. Most people find the backward game harder. That raises a question: does a reversed text carry more information per character than the text read forward (MacKay 2003, exercise 8.11)?

Exercise 0.2 (The whole text costs the same in both orders)**.** Let $X_1, \dots, X_n$ be discrete random variables. Use the chain rule for entropy, with the variables listed in each of two orders, to show that

$$\sum_{t=1}^{n} \mathbb{H}(X_t \mid X_1, \dots, X_{t-1}) = \sum_{t=1}^{n} \mathbb{H}(X_t \mid X_{t+1}, \dots, X_n),$$

where the term for $t = 1$ on the left and the term for $t = n$ on the right are unconditional entropies.

Solution

Solution 0.2. Listing the same variables in another order does not change their joint entropy, because the reordered vector takes each reordered value with the same probability, as in the proof of Corollary 0.2. For the same reason, reordering the variables in a condition does not change a conditional entropy. The chain rule, with the variables in their usual order, gives

$$\mathbb{H}(X_1, \dots, X_n) = \sum_{t=1}^{n} \mathbb{H}(X_t \mid X_1, \dots, X_{t-1}).$$

With the variables listed as $X_n, X_{n-1}, \dots, X_1$, the ith variable is X_{n+1-i}, and the variables before it are $X_n, \dots, X_{n+2-i}$. So

$$\begin{aligned}
\mathbb{H}(X_1, \dots, X_n) &= \mathbb{H}(X_n, \dots, X_1) \\
&= \sum_{i=1}^{n} \mathbb{H}(X_{n+1-i} \mid X_n, \dots, X_{n+2-i}) \\
&= \sum_{t=1}^{n} \mathbb{H}(X_t \mid X_n, \dots, X_{t+1}) \\
&= \sum_{t=1}^{n} \mathbb{H}(X_t \mid X_{t+1}, \dots, X_n).
\end{aligned}$$

The first line reorders the joint entropy, the second is the chain rule, the third substitutes $t = n + 1 - i$, and the fourth reorders each condition. Both sums equal $\mathbb{H}(X_1, \dots, X_n)$, so they equal each other.

Exercise 0.3 (Conditional entropy of an N-gram model read forward)**.** Let $X_1, X_2, \dots$ be a stationary sequence that is also an N-gram model with $N \geq 2$. For every $t > N$, show that the forward conditional entropy of X_t is

$$\mathbb{H}(X_t \mid X_1, \dots, X_{t-1}) = \mathbb{H}(X_1, \dots, X_N) - \mathbb{H}(X_1, \dots, X_{N-1}).$$

Solution

Solution 0.3. To shorten the displays, write $p(x_a, \dots, x_b)$ for $\mathbb{P}(X_a = x_a, \dots, X_b = x_b)$, and $p(x_a \mid x_c, \dots, x_d)$ for $\mathbb{P}(X_a = x_a \mid X_c = x_c, \dots, X_d = x_d)$. Each value's subscript says which variable takes it.

For values $x_1, \dots, x_{t-1}$ with $p(x_1, \dots, x_{t-1}) > 0$, Definition 0.9 makes $\mathbb{H}(X_t \mid X_1 = x_1, \dots, X_{t-1} = x_{t-1})$ a function of the conditional probability mass function $p(\cdot \mid x_1, \dots, x_{t-1})$ alone. The N-gram model makes that conditional probability mass function equal to $p(\cdot \mid x_{t-N+1}, \dots, x_{t-1})$, so

$$\mathbb{H}(X_t \mid X_1 = x_1, \dots, X_{t-1} = x_{t-1}) = \mathbb{H}(X_t \mid X_{t-N+1} = x_{t-N+1}, \dots, X_{t-1} = x_{t-1}).$$

Then

$$\begin{aligned}
\mathbb{H}(X_t \mid X_1, \dots, X_{t-1}) &= \sum_{x_1,\dots,x_{t-1}} p(x_1, \dots, x_{t-1})\, \mathbb{H}(X_t \mid X_1 = x_1, \dots, X_{t-1} = x_{t-1}) \\
&= \sum_{x_1,\dots,x_{t-1}} p(x_1, \dots, x_{t-1})\, \mathbb{H}(X_t \mid X_{t-N+1} = x_{t-N+1}, \dots, X_{t-1} = x_{t-1}) \\
&= \sum_{x_{t-N+1},\dots,x_{t-1}} p(x_{t-N+1}, \dots, x_{t-1})\, \mathbb{H}(X_t \mid X_{t-N+1} = x_{t-N+1}, \dots, X_{t-1} = x_{t-1}) \\
&= \mathbb{H}(X_t \mid X_{t-N+1}, \dots, X_{t-1}).
\end{aligned}$$

The first line is Theorem 0.7, with $(X_1, \dots, X_{t-1})$ as the conditioning variable. The second line uses the display above. The third line sums over $x_1, \dots, x_{t-N}$ first: the second factor of each term does not depend on them, and summing $p(x_1, \dots, x_{t-1})$ over them gives the marginal $p(x_{t-N+1}, \dots, x_{t-1})$. The fourth line is Theorem 0.7 again.
The two-variable chain rule, with $(X_{t-N+1}, \dots, X_{t-1})$ as its first variable and X_t as its second, gives

$$\mathbb{H}(X_{t-N+1}, \dots, X_t) = \mathbb{H}(X_{t-N+1}, \dots, X_{t-1}) + \mathbb{H}(X_t \mid X_{t-N+1}, \dots, X_{t-1}).$$

Subtracting $\mathbb{H}(X_{t-N+1}, \dots, X_{t-1})$ from both sides gives

$$\mathbb{H}(X_t \mid X_{t-N+1}, \dots, X_{t-1}) = \mathbb{H}(X_{t-N+1}, \dots, X_t) - \mathbb{H}(X_{t-N+1}, \dots, X_{t-1}).$$

An entropy depends only on the joint probability mass function. The stationary sequence with shift $s = t - N$ gives $(X_{t-N+1}, \dots, X_t)$ the same joint probability mass function as $(X_1, \dots, X_N)$, and $(X_{t-N+1}, \dots, X_{t-1})$ the same one as $(X_1, \dots, X_{N-1})$. So

$$\mathbb{H}(X_t \mid X_1, \dots, X_{t-1}) = \mathbb{H}(X_1, \dots, X_N) - \mathbb{H}(X_1, \dots, X_{N-1}).$$

Exercise 0.4 (An N-gram model read backward)**.** Let $X_1, X_2, \dots$ be an N-gram model with $N \geq 2$, let $n \geq N$, and let $1 \leq t \leq n - N + 1$. Show that given all the later variables, X_t depends only on the $N - 1$ variables just after it: for every choice of values $x_t, \dots, x_n$ with $\mathbb{P}(X_{t+1} = x_{t+1}, \dots, X_n = x_n) > 0$,

$$\mathbb{P}(X_t = x_t \mid X_{t+1} = x_{t+1}, \dots, X_n = x_n) = \mathbb{P}(X_t = x_t \mid X_{t+1} = x_{t+1}, \dots, X_{t+N-1} = x_{t+N-1}).$$

Solution

Solution 0.4. To shorten the displays, write $p(x_a, \dots, x_b)$ for $\mathbb{P}(X_a = x_a, \dots, X_b = x_b)$, and $p(x_a \mid x_c, \dots, x_d)$ for $\mathbb{P}(X_a = x_a \mid X_c = x_c, \dots, X_d = x_d)$. Each value's subscript says which variable takes it.

First, for each s with $t + N \le s \le n$, the condition on $X_t, \dots, X_{s-1}$ can be cut to the $N - 1$ variables just before X_s: for values with $p(x_t, \dots, x_{s-1}) > 0$,

$$\begin{aligned} p(x_s \mid x_t, \dots, x_{s-1}) &= \sum_{x_1, \dots, x_{t-1}} p(x_1, \dots, x_{t-1} \mid x_t, \dots, x_{s-1})\, p(x_s \mid x_1, \dots, x_{s-1}) \\ &= \sum_{x_1, \dots, x_{t-1}} p(x_1, \dots, x_{t-1} \mid x_t, \dots, x_{s-1})\, p(x_s \mid x_{s-N+1}, \dots, x_{s-1}) \\ &= p(x_s \mid x_{s-N+1}, \dots, x_{s-1}) \sum_{x_1, \dots, x_{t-1}} p(x_1, \dots, x_{t-1} \mid x_t, \dots, x_{s-1}) \\ &= p(x_s \mid x_{s-N+1}, \dots, x_{s-1}). \end{aligned}$$

The first line is the law of total probability, given $X_t, \dots, X_{s-1}$. The second line is the N-gram model, which applies because $s > N$. The third line takes the factor that does not depend on $x_1, \dots, x_{t-1}$ out of the sum. The fourth line uses that a conditional probability mass function sums to 1. Because $s - N + 1 \ge t + 1$, the same holds with $t + 1$ in place of t.

Write R for the product of these factors:

$$R = \prod_{s=t+N}^{n} p(x_s \mid x_{s-N+1}, \dots, x_{s-1}).$$

The product is empty, and R is 1, when $n = t + N - 1$. No factor involves x_t, because each starts at x_{s-N+1} with $s - N + 1 \ge t + 1$. Applying the multiplication rule once for each s from $t + N$ to n, and then the cut above to each factor, gives

$$\begin{aligned} p(x_t, \dots, x_n) &= p(x_t, \dots, x_{t+N-1}) \prod_{s=t+N}^{n} p(x_s \mid x_t, \dots, x_{s-1}) \\ &= p(x_t, \dots, x_{t+N-1})\, R. \end{aligned}$$

The same two steps, starting from x_{t+1}, give

$$\begin{aligned} p(x_{t+1}, \dots, x_n) &= p(x_{t+1}, \dots, x_{t+N-1}) \prod_{s=t+N}^{n} p(x_s \mid x_{t+1}, \dots, x_{s-1}) \\ &= p(x_{t+1}, \dots, x_{t+N-1})\, R. \end{aligned}$$

The left side of the second display is positive by assumption, so $R > 0$ and $p(x_{t+1}, \dots, x_{t+N-1}) > 0$. By the definition of conditional probability,

$$
\begin{aligned}
p(x_t \mid x_{t+1}, \dots, x_n) &= \frac{p(x_t, \dots, x_n)}{p(x_{t+1}, \dots, x_n)} \\
&= \frac{p(x_t, \dots, x_{t+N-1})\, R}{p(x_{t+1}, \dots, x_{t+N-1})\, R} \\
&= \frac{p(x_t, \dots, x_{t+N-1})}{p(x_{t+1}, \dots, x_{t+N-1})} \\
&= p(x_t \mid x_{t+1}, \dots, x_{t+N-1}).
\end{aligned}
$$

The second line substitutes the two displays above, the third cancels R, and the fourth is the definition of conditional probability again.

Exercise 0.5 (Conditional entropy of an N-gram model read backward)**.** Let $X_1, X_2, \dots$ be a stationary sequence that is also an N-gram model with $N \geq 2$, let $n \geq N$, and let $1 \leq t \leq n - N + 1$. Use Exercise 0.4 to show that the backward conditional entropy of X_t is

$$
\mathbb{H}(X_t \mid X_{t+1}, \dots, X_n) = \mathbb{H}(X_1, \dots, X_N) - \mathbb{H}(X_1, \dots, X_{N-1}).
$$

Solution

Solution 0.5. To shorten the displays, write $p(x_a, \dots, x_b)$ for $\mathbb{P}(X_a = x_a, \dots, X_b = x_b)$. Each value's subscript says which variable takes it.

The reasoning of Solution 0.3 carries over, with the later variables in place of the earlier ones and Exercise 0.4 in place of the N-gram model:

$$
\begin{aligned}
\mathbb{H}(X_t \mid X_{t+1}, \dots, X_n) &= \sum_{x_{t+1}, \dots, x_n} p(x_{t+1}, \dots, x_n)\, \mathbb{H}(X_t \mid X_{t+1} = x_{t+1}, \dots, X_n = x_n) \\
&= \sum_{x_{t+1}, \dots, x_n} p(x_{t+1}, \dots, x_n)\, \mathbb{H}(X_t \mid X_{t+1} = x_{t+1}, \dots, X_{t+N-1} = x_{t+N-1}) \\
&= \sum_{x_{t+1}, \dots, x_{t+N-1}} p(x_{t+1}, \dots, x_{t+N-1})\, \mathbb{H}(X_t \mid X_{t+1} = x_{t+1}, \dots, X_{t+N-1} = x_{t+N-1}) \\
&= \mathbb{H}(X_t \mid X_{t+1}, \dots, X_{t+N-1}).
\end{aligned}
$$

The first and fourth lines are Theorem 0.7. The second line uses Exercise 0.4 and Definition 0.9. The third line sums over $x_{t+N}, \dots, x_n$ first, which gives the marginal $p(x_{t+1}, \dots, x_{t+N-1})$.

The chain rule, with $(X_{t+1}, \dots, X_{t+N-1})$ as its first variable, then gives

$$
\mathbb{H}(X_t \mid X_{t+1}, \dots, X_{t+N-1}) = \mathbb{H}(X_t, \dots, X_{t+N-1}) - \mathbb{H}(X_{t+1}, \dots, X_{t+N-1}).
$$

The stationary sequence with shift $s = t - 1$ gives $(X_t, \dots, X_{t+N-1})$ the same joint proba-

bility mass function as $(X_1, \dots, X_N)$, and with shift $s = t$ it gives $(X_{t+1}, \dots, X_{t+N-1})$ the same one as $(X_1, \dots, X_{N-1})$. So

$$\mathbb{H}(X_t \mid X_{t+1}, \dots, X_n) = \mathbb{H}(X_1, \dots, X_N) - \mathbb{H}(X_1, \dots, X_{N-1}).$$

Exercise 0.6 (Which guessing game is harder?)**.** Use Exercise 0.2, Exercise 0.3 and Exercise 0.5 to compare the information per character in a reversed text and in the text read forward, and use that comparison to explain the difference in difficulty described in Remark 0.17.

Solution

Solution 0.6. Exercise 0.3 and Exercise 0.5 give each character away from the ends of the text the same conditional entropy forward and backward, and Exercise 0.2 gives the whole text the same sum of conditional entropies forward and backward. So the reversed text carries no more information per character than the text read forward. The two games tie only on average: in English, "q" is almost always followed by "u", but "u" is rarely preceded by "q", so a single guess can be easy one way and hard the other.
People find the backward game harder because their guesses are worse, not because the text holds more information. A reader who guesses with a probability mass function g in place of the true conditional probability mass function f spends $\mathbb{H}(f, g)$ bits per character on average, which exceeds $\mathbb{H}(f)$ by $D_{\mathrm{KL}}(f \,\|\, g)$, by the decomposition of cross-entropy. Years of reading forward make g close to f for forward guessing, and no such practice makes it close backward.

References

Cover, Thomas M., and Joy A. Thomas. 2006. *Elements of Information Theory.* 2nd ed. Wiley-Interscience. https://doi.org/10.1002/047174882X.

Grinstead, Charles M., and J. Laurie Snell. 1997. *Introduction to Probability.* 2nd ed. American Mathematical Society.

International Electrotechnical Commission. 2008. *Quantities and Units — Part 13: Information Science and Technology.* IEC 80000-13:2008. International Electrotechnical Commission.

Jelinek, F., R. L. Mercer, L. R. Bahl, and J. K. Baker. 1977. "Perplexity—a Measure of the Difficulty of Speech Recognition Tasks." *The Journal of the Acoustical Society of America* 62 (S1): S63. https://doi.org/10.1121/1.2016299.

Leinster, Tom. 2021. *Entropy and Diversity: The Axiomatic Approach*. Cambridge University Press. https://doi.org/10.1017/9781108963558.

MacKay, David J. C. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning* 1 (1): 81–106. https://doi.org/10.1007/BF00116251.

Rényi, Alfréd. 1961. "On Measures of Entropy and Information." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, edited by Jerzy Neyman. University of California Press. https://projecteuclid.org/euclid.bsmsp/1200512181.

Wikipedia contributors. 2026. *Rényi Entropy — Wikipedia, the Free Encyclopedia*. https://en.wikipedia.org/wiki/R%C3%A9nyi_entropy.