Introduction to Entropy

Erik Stern

December 8, 2025

Abstract

Your abstract.

Contents

1 Entropy and Mutual information

1.1 Definitions and Conventions

Definition 1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{X}, \mathcal{Y}$ be countable sets and $X : \Omega \to \mathcal{X}$, $Y : \Omega \to \mathcal{Y}$ be discrete random variables on $(\Omega, \mathcal{A}, \mathbb{P})$.
We can then define

$$\text{Entropy wrt. base:} \quad H_q(X) = \mathbb{E}\left(-\log_q p(X)\right) = -\sum_{x \in \operatorname{supp}(X)} p(x) \log_q p(x)$$

$$\text{Entropy conventionally:} \quad H(X) = H_2(X)$$

Remark 1.
1. We use the shorthand notations $p(x) = \mathbb{P}[X = x]$ and $p(y) = \mathbb{P}[Y = y]$. If the context is unclear, the normal notations will be used to avoid confusion!
2. Currently, $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$ is not required. Later theorems like Jensens Inequality do require real-valued random variables.
3. We use the convention $\log = \log_2$, as the entropy H is defined wrt. base 2.
4. We also use the following convention and justify it through a continuity argument:

$$0 \log 0 = \lim_{x \to 0^+} x \log x = \lim_{x \to 0^+} \frac{\log x}{\frac{1}{x}} = \lim_{x \to 0^+} \frac{\frac{1}{\ln(2)x}}{-\frac{1}{x^2}} = \lim_{x \to 0^+} \frac{-x}{\ln(2)} = 0$$

   This choice is sensible, as $\log x$ is not defined for negative x.
5. The conventions, definitions and theorems are from Definitions, Theorems, Remarks and Exercises in *Elements of Information Theory, second edition* (see [CT05]).

Remark 2. Note that if $|\mathcal{X}|$ is finite, $(\forall p \in \mathbb{R}_+ : H_p(X) \text{ finite})$ and $H(X) \leq |\mathcal{X}|$ (see Theorem 4). For $|\mathcal{X}|$ countably infinite, there are counterexamples where $H_p(X) = \infty$ (see [Mat]). From now on, we will assume that entropy is finite.

(a) Bernoulli Rand. Variable Entropy $H(X_p)$ (b) Geometric Rand. Variable Entropy $H(X_p)$

Example 1. Let $p \in (0,1)$ and $X_p \sim B(1,p)$ be a weighted coin flip.
We can calculate the Entropy of X_p: $H(X_p) = -p \log p - (1-p)\log(1-p)$.
A visual inspection (see Figure 1a) reveals that $H(X_p)$ seems to be maximised for p and minimised for $p \in \{0,1\}$. An increase in uncertainty about the result of the coin flip seems to correspond with an increase in entropy.

Example 2. Let $p \in (0,1)$ and $X_p \sim G(p)$ be the number of times a weighted coin is flipped, until the first head occurs. We will calculate the Entropy of X_p. It will require the two well-known series:

$$\forall r \in (0,1) : \sum_{n \in \mathbb{N}_0} r^n = \frac{1}{1-r} \tag{1}$$

$$\forall r \in (0,1) : \sum_{n \in \mathbb{N}_0} n r^n = \frac{r}{(1-r)^2} \tag{2}$$

We can now directly calculate the Entropy of X_p:

$$\begin{aligned}
H(X_p) &= \sum_{x \in \mathbb{N}} -p(x) \log p(x) && \text{(Def. of Entropy)} \\
&= -\sum_{x \in \mathbb{N}} (1-p)^{x-1} p \log\left((1-p)^{x-1} p\right) && \text{(Subst. in geometric density function)} \\
&= -\sum_{x \in \mathbb{N}} (1-p)^{x-1} p \left((x-1)\log(1-p) + \log p\right) && \text{(Log rules)} \\
&= -p\log(1-p) \sum_{x \in \mathbb{N}_0} ((1-p)^x x) - p \log p \sum_{x \in \mathbb{N}_0} ((1-p)^x) && \text{(Factor out constants)} \\
&= -p\log(1-p) \frac{1-p}{(1-(1-p))^2} - p \log p \frac{1}{1-(1-p)} && \text{(Use series 1 and 2)} \\
&= -p\log(1-p) \frac{1-p}{p^2} - p \log p \frac{1}{p} && \text{(Simplify expr.)} \\
&= \frac{-(1-p)\log(1-p) - p \log p}{p} && \text{(Simplify expr.)}
\end{aligned}$$

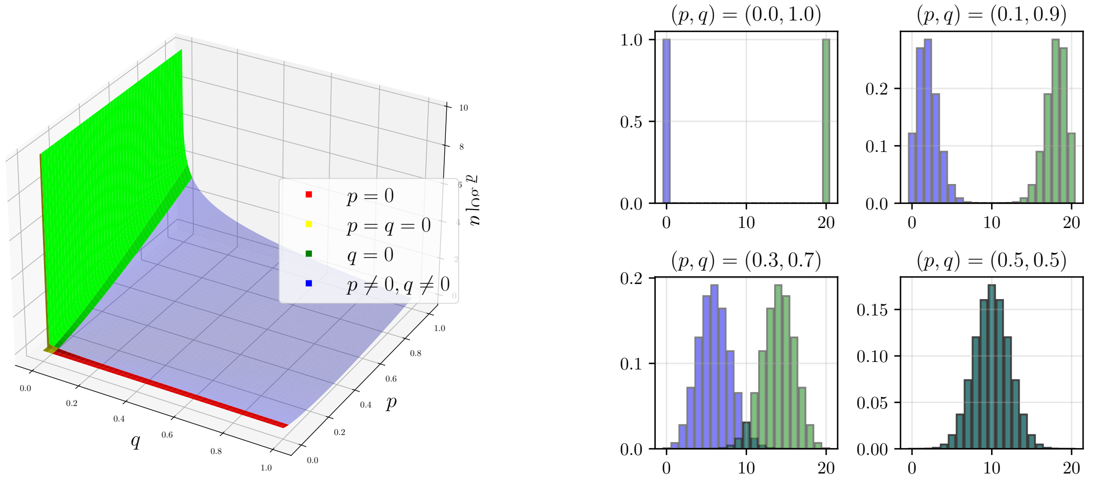
For $p = 0.5$ we get $H(X_{0.5}) = \frac{-0.5 \log 0.5 - 0.5 \log 0.5}{0.5} = -2 \log 0.5 = 2$.
We can visually inspect $(0,1) \to \mathbb{R}, p \mapsto H(X_p)$ (see Figure 1b) to get a feeling for the entropy of X_p. An increase in p, so less variance and concentration towards zero, seems to be linked to a lower entropy and vice-versa.

TODO: Complete the q

Definition 2. Let X, Y be discrete random variables with marginal distributions $p(x), p(y)$ and with joint distribution $p(x, y)$. We define the following:

$$\begin{aligned}
\text{Conditional Entropy:} \quad & H(X \mid Y) = -\mathbb{E}\left(\log p(X \mid Y)\right) \\
\text{Joint Entropy:} \quad & H(X, Y) = -\mathbb{E}\left(\log p(X, Y)\right) \\
\text{Relative Entropy:} \quad & D(p(x) \| q(x)) = \mathbb{E}_p\left(\log \frac{p(X)}{q(X)}\right) \\
& \text{with the conventions} 0 \log 0 = 0, \\
\text{Mutual information:} \quad & I(X; Y) = D(p(x, y) \| p(x)p(y)) \\
\text{Conditional Mutual information:} \quad & I(X; Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)
\end{aligned}$$

Remark 3. From now on, we always define the two discrete random variables X, Y from Definition 2.



(a) Pointwise Relative Entropy

(b) Relative Entropies of Binomial Distributions

Remark 4. Under the assumptions of Definition 2, we have

$$\begin{aligned} D(p(x) \| q(x)) &= \mathbb{E}_p \left(\log \frac{p(X)}{q(X)} \right) && \text{(def. of relative entropy)} \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} && \text{(def. of expected value)} \end{aligned}$$

To understand the conventions, we can look at the limit cases:

1. **Case** $p \in (0, 1], q = 0$**:** $\lim_{q \to 0^+} p \log \frac{p}{q} = \lim_{q \to 0^+} (p \log p - p \log q) = \infty$.
2. **Case** $p = 0, q \in (0, 1]$**:** $0 \log \frac{0}{q} = 0$.
3. **Case** $p = q = 0$**:** Case 1 logic yields $\lim_{q \to 0^+} p \log \frac{p}{q} = \infty$ and Case 2 logic yields $0 \log \frac{0}{0} = 0$.

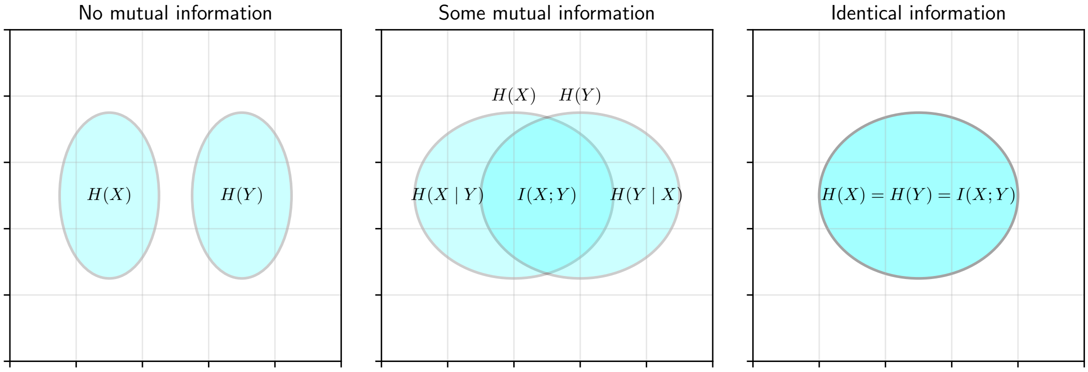
As we want $\sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$ to sum over $x \in \mathcal{X}, p(x) > 0$, we choose the convention $0 \log \frac{0}{0} = 0$. Figure 2a visualizes the the pointwise relative entropy function $(p, q) \mapsto \log \frac{p}{q}$: We can now calculate the relative entropies for an example. Let $X \sim B(20, \alpha)$ and $Y \sim B(20, \beta)$ with $(\alpha, \beta) \in [0, 1]^2$.

$$\begin{aligned} & D(p(x) \| q(x)) = \sum_{x=0}^{20} p(x) \log \frac{p(x)}{q(x)} \\ \alpha = 0, \beta = 1: \quad & D(p(x) \| q(x)) = 0 \log \frac{0}{1} + 1 \log \frac{1}{0} = 0 + \infty = \infty \\ \alpha = 0.1, \beta = 0.9: \quad & D(p(x) \| q(x)) \approx 50.7 \\ \alpha = 0.3, \beta = 0.7: \quad & D(p(x) \| q(x)) \approx 9.8 \\ \alpha = 0.5, \beta = 0.5: \quad & D(p(x) \| q(x)) = \sum_{x=0}^{20} p(x) \log 1 = 0 \end{aligned}$$

Figure 2b visualizes the two discrete distribution functions in the cases above. Intuitively, the more overlap the distributions have, the closer to zero the relative entropy is.

1.2 Chain Rules and Mutual Information

Remark 5. The relationship between entropy, conditional entropy and mutual information can be visualized:



Theorem 1. *Let $n \in \mathbb{N}, n \geq 2$, $(X_1, \cdots, X_n) \sim p(x_1, \cdots, x_n)$. The following statements about Entropy and Mutual Information are called Chain Rules:*

1. $H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \cdots, X_1)$
2.
3.

Proof. 1. Prove this result using induction by n.
Base case $n = 2$:

$$\begin{aligned} H(X_1, X_2) &= -\mathbb{E}\left(\log p(X_1, X_2)\right) \\ &= -\mathbb{E}\left(\log p(X_2 \mid X_1) p(X_1)\right) \\ &= -\mathbb{E}\left(\log p(X_2 \mid X_1)\right) + -\mathbb{E}\left(\log p(X_1)\right) \\ &= H(X_1) + H(X_2 \mid X_1) \end{aligned}$$

Assume the theorem holds for $n-1$. Induction case $n-1$ to n:

$$\begin{aligned} H(X_1, \cdots, H_n) &= H(X_n | X_1, \cdots, H_{n-1}) + H(X_1, \cdots, H_{n-1}) && \text{(apply base case)} \\ &= H(X_n | X_1, \cdots, H_{n-1}) + \sum_{i=1}^{n-1} H(X_i \mid X_{i-1}, \cdots, X_1) && \text{(induction hypothesis)} \end{aligned}$$

□

Corollary 1. *We can specialize the Chain Rule for Entropy and Mutual Information to two variables:*

1. $H(X, Y) = H(X) + H(Y \mid X)$
2. $I(X;Y) =$

Proof. Follows from Theorem 1. □

Theorem 2. *There are multiple equivalent ways to express Mutual Information:*

1. $I(X;Y) = H(Y) - H(Y|X)$
2. $I(X;Y) = I(Y;X)$
3. $I(Y;X) = H(X) - H(X|Y)$
4. $I(X;Y) = H(X,Y)$
5. $I(X;X) = H(X)$

Proof. 1. We can use the definition of mutual information and relative entropy to obtain:

$$\begin{aligned}
I(X;Y) &= D(p(x,y) \| p(x)p(y)) && \text{(by def. of mutual info.)} \\
&= \mathbb{E}_{p(x,y)} \left(\log \frac{p(x,y)}{p(x)p(y)} \right) && \text{(by def. relative entropy)} \\
&= \mathbb{E}_{p(x,y)} \left(\log \frac{p(x)p(y|x)}{p(x)p(y)} \right) && \text{(using cond. probability)} \\
&= \mathbb{E}_{p(x,y)} \left(\log \frac{p(y|x)}{p(y)} \right) && \text{(simplify fraction)} \\
&= \mathbb{E}_{p(x,y)} \left(\log p(y|x) \right) - \mathbb{E}_p(x,y) \left(\log p(y) \right) && \text{(simplify logarithm)} \\
&= H(Y|X) - H(Y) && \text{(by def. of entropy)}
\end{aligned}$$

2. The definition of mutual information yields:

$$\begin{aligned}
I(X;Y) &= D(p(x,y) \| p(x)p(y)) && \text{(by def. of mutual info.)} \\
&= D(P[X=x, Y=y] \| P[X=x]P[Y=y]) && \text{(use clear notation)} \\
&= D(P[Y=y, X=x] \| P[Y=y]P[X=x]) && \text{(use commutativity)} \\
&= D(p(y,x) \| p(y)p(x)) && \text{(back to conventional notation)} \\
&= I(X;Y)
\end{aligned}$$

3. Follows directly from 2 and 3.

4.

$$\begin{aligned}
I(X;Y) &= H(Y) - H(X|Y) && \text{(by 1)} \\
&= H(Y) - (H(X,Y) - H(X)) && \text{(chain rule)} \\
&= H(X) + H(Y) - H(X,Y)
\end{aligned}$$

5. Using 1 we get $I(X;X) = H(X) - H(X|X) = H(X)$.

□

2 Inequalities for Entropy and Mutual Information

2.1 Convexity and Jensens Inequality

Remark 6. We use the common definition of convex functions, concave functions from analysis. From now on, X, Y have to be real-valued random variables.

Theorem 3. *Let $f : \Omega \to \mathbb{R}$ a function.*

1. *If f is convex, we have $\mathbb{E}f(X) \geq f(\mathbb{E}X)$.*
2. *If f is concave, we have $\mathbb{E}f(X) \leq f(\mathbb{E}X)$.*

Proof. 1. Distinguish two cases. If $\mathcal{X}$ is finite:
We show $f\left(\sum_{i=1}^{n} p_i x_i\right) \leq \sum_{i=1}^{n} p_i f(x_i)$ by induction.
The definition of convexity yields the base case $i = 2$: $f(p_1 x_1 + p_2 x_2) \leq p_1 f(x_1) + p_2 f(x_2)$.
We assume the claim holds for $n - 1$ and the induction case goes as follows:

$$\begin{aligned}
f\left(\sum_{i=1}^{n} p_i x_i\right) &= f\left(p_1 x_1 + (1 - p_1) \sum_{i=2}^{n} \frac{p_i}{1 - p_1} x_i\right) \\
&\leq p_1 f(x_1) + (1 - p_1) f\left(\sum_{i=2}^{n} \frac{p_i}{1 - p_1} x_i\right) && \text{(def. of convexity)} \\
&\leq p_1 f(x_1) + (1 - p_1) \sum_{i=2}^{n} \frac{p_i}{1 - p_1} f(x_i) && \text{(induct. hypo. applies bc. of } \sum_{i=2}^{n} \frac{p_i}{1 - p_1} = 1) \\
&= \sum_{i=1}^{n} p_i f(x_i)
\end{aligned}$$

Else:

2. Follows from part 1 applied to $-f$.

$\square$

Corollary 2. *Entropy and Mutual Information are non-negative:*

1. $0 \leq H(X)$
2. $0 \leq D(p(x) \| q(x))$
3. $0 \leq I(X; Y)$

Proof. 1. Note that $\log(\frac{1}{[0,1]}) = \log([1, \infty]) = [0, \infty]$ and $p(X)(\mathcal{X}) \in [0, 1]$.
Using the monotonicity of the expected value, we obtain

$$0 \leq \mathbb{E}\left(\log\left(\frac{1}{p(X)}\right)\right) = -\mathbb{E}\left(\log\left(p(X)\right)\right) = H(X)$$

2. We can prove this using Jensens Inequality on a *concave* function:

$$\begin{aligned}
-D(p(x) \| q(x)) &= -\mathbb{E}_p\left(\log \frac{p(X)}{q(X)}\right) && \text{(def. of relative entropy)} \\
&\leq -\log\left(\mathbb{E}_p \frac{p(X)}{q(X)}\right) = \log\left(\mathbb{E}_p \frac{q(X)}{p(X)}\right) && \text{(-log is convex, property of log)} \\
&= \log\left(\sum_{x \in \mathcal{X}} p(x) \frac{q(x)}{p(x)}\right) = \log\left(\sum_{x \in \mathcal{X}} q(x)\right) && \text{(def. of exp. value, simplify expr.)} \\
&= \log(1) = 0 && \text{(q is a prob. function)}
\end{aligned}$$

So equivalently, we have $D(p(x) \| q(x)) \geq 0$.

3. Follows from part 1: $I(X;Y) = D(p(x,y) \| p(x)p(y)) \geq 0$.

$\square$

Remark 7. There are multiple natural questions we can ask about Entropy. We will look at an example for each of them and then prove the results in the follow-up theorem.

1. What distribution maximises the value of Entropy? I simulated

   From the data, it seems plausible that the uniform distribution maximises entropy. We

2. Can joint entropy increase if we add redundant information?
   Let $\Omega = \{1, 2\}$, p uniform TODO, $X, Y : \Omega \to \mathbb{R}$, $X(\omega) = \omega$ and $Y(\omega) = 2\omega$. Y is redundant to X, as $Y = 2X$.

$$H(X) = \sum_{i=1}^{2} -0.5 \log 0.5 = -\log 0.5 = \log 2 = 1$$
$$H(X,Y) = \sum_{(i,j) \in \{(1,2),(2,4)\}} -0.25 \log 0.25 = 2 * 0.25 * 2 = 1$$

   Pairs like $(1,4)$ have probability zero and do not contribute to the sum in $H(X,Y)$. This example suggest entropy never increases. We will prove that in the follow-up theorem.

3. What happens to the Entropy if we add independent noise to our measurements?
   Let $X \sim U(\{1,2,3,4\})$ be the original signal, $N \sim B(3, 0.5)$ the noise and let X, N be independent. Then $S := X + N$ is a noisy signal.

   We can compute the crirical variance a.

Theorem 4. *We can formalise the previous observations including some more:*

1. *More information can only decrease entropy:* $H(X \mid Y) \leq H(X)$.

2. *The uniform distribution maximizes entropy:*
   $H(X) \leq \log |\mathcal{X}|$ *and* $H(X) = \log |\mathcal{X}| \iff X \sim U(\mathcal{X})$.

3. *If information from* Y *does not add anything to* X*, then* Y *must be derived from* X*:*
   $H(Y \mid X) = 0 \implies \exists f : Y = f(X)$ *almost surely*

4. *If independent noise is added to a random variable, entropy can only increase:*
   Set $Z = X + Y$*. Then we have* X, Y *independent* $\implies H(X) \leq H(Z) \wedge H(Y) \leq H(Z)$

Proof.

1. $0 \leq I(X;Y) = H(X) - H(X \mid Y) \iff H(X \mid Y) \leq H(X)$

2. Let $Y \sim U(\mathcal{X})$ st. $\forall x \in \mathcal{X} : q(x) = \frac{1}{|\mathcal{X}|}$.

$$0 \leq D(p(x) \| q(x)) = \mathbb{E}_p \left(\log \frac{p(X)}{q(X)} \right) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$
$$= \sum_{x \in \mathcal{X}} p(x) \log \left(p(x) |\mathcal{X}| \right) = \sum_{x \in \mathcal{X}} p(x) \log p(x) + |\mathcal{X}| \sum_{x \in \mathcal{X}} p(x)$$
$$= -H(X) + |\mathcal{X}| = |\mathcal{X}| - H(X)$$

   This is equivalent to $H(X) \leq \log |\mathcal{X}|$. TODO: Iff part.

3. We have $H(Y \mid X) = -\mathbb{E}(\log p(Y \mid X)) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} -p(x,y) \log p(y \mid x)$.
   Additionally, we have $\forall (x,y) \in \mathcal{X} \times \mathcal{Y} : -p(x,y) \log p(y \mid x) \geq 0$.
   Combining those facts, we get

$$H(Y \mid X) = 0$$
$$\iff \forall (x,y) \in \mathcal{X} \times \mathcal{Y} : p(x,y) \log p(y \mid x) = 0$$
$$\iff \forall (x,y) \in \mathcal{X} \times \mathcal{Y} : p(x,y) = 0 \oplus \log p(y \mid x) = 0$$
$$\iff \forall (x,y) \in \mathcal{X} \times \mathcal{Y} : p(x,y) = 0 \oplus p(y \mid x) = 1$$

This tells us, that either $p(x,y) = 0$ or $p(x,y) = p(y \mid x)p(x) = p(x) = 1$.
Now we can finish up the argument. Set $A = \{x \in \mathcal{X} : p(x) > 0\}$.
Define $(y_x)_{x\in\mathcal{X}}$ such that $\forall x \in \mathcal{X} : p(x, y_x) > 0$.
Set $f : A \to \mathcal{Y}, x \mapsto y_x$. This gets us $Imf = \{y_x : x \in \mathcal{X}\} = \{y \in \mathcal{Y} : p(x,y) > 0\}$.
So $Y = f(X)$ almost surely.

4. We have

$$
\begin{aligned}
H(Z \mid X) = -\mathbb{E}\left(\log p(Z \mid X)\right) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} -p(x,y)\log P[Z = x + y \mid X = x] \\
= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} -p(x,y)\log P[X + Y = x + y \mid X = x] \\
= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} -p(x,y)\log P[Y = y \mid X = x] \\
= -\mathbb{E}\left(\log p(Y \mid X)\right) = H(Y \mid X)
\end{aligned}
$$

$H(Z,X) = H(X) + H(Z \mid X)$ $H(Y,X) = H(X) + H(Y \mid X)$
$H(Y,X) = H(X) + H(Y \mid X)$ Using the Chain rule, we get $H(X,Y) = H(Y \mid X) + H(X)$
□

2.2 The Information Inequality

Hello

2.3 Application to Optimisation

Hello

References

[CT05] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 1st ed. Wiley, Sept. 16, 2005. ISBN: 9780471241959 9780471748823. DOI: 10.1002/047174882X. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X (visited on 12/08/2025).

[Mat] *Can the entropy of a random variable with countably many outcomes be infinite?* URL: https://math.stackexchange.com/questions/279304/can-the-entropy-of-a-random-variable-with-countably-many-outcomes-be-infinite (visited on 08/12/2025).