Channel capacity

Last modified: 2026-11-18 21:21:45 (PST)

A channel carries symbols from a sender to a receiver, and some of them are corrupted along the way. This page asks how much information a channel can carry, using the mutual information between what goes in and what comes out.

Channels

Definition 0.1 (Discrete memoryless channel)**.** A **discrete channel** has a finite *input alphabet* $\mathcal{X}$, a finite *output alphabet* $\mathcal{Y}$, and, for each input $x \in \mathcal{X}$, a probability mass function $p(y \mid x)$ on $\mathcal{Y}$: the probability that the channel outputs y when it is sent x. The channel is **memoryless** if, over n uses, each output depends only on the input sent at that use:

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n \mid X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} p(y_i \mid x_i).$$

This is the setup of Cover and Thomas (2006, chap. 7).

Definition 0.2 (Binary symmetric channel)**.** The **binary symmetric channel** with *crossover probability* $\epsilon \in [0, 1]$ is the discrete memoryless channel with $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ that flips its input with probability ϵ:

$$p(y \mid x) = \begin{cases} 1 - \epsilon & \text{if } y = x, \\ \epsilon & \text{if } y \neq x. \end{cases}$$

Definition 0.3 (Binary erasure channel)**.** The **binary erasure channel** with *erasure probability* $\alpha \in [0, 1]$ is the discrete memoryless channel with $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, \mathrm{e}, 1\}$ that delivers its input with probability $1 - \alpha$ and otherwise outputs the erasure symbol e:

$$p(y \mid x) = \begin{cases} 1-\alpha & \text{if } y = x, \\ \alpha & \text{if } y = \mathrm{e}, \\ 0 & \text{otherwise.} \end{cases}$$

Remark 0.1 (Erasures are visible). On the binary erasure channel the receiver knows which symbols were lost, but not what they were. On the binary symmetric channel the receiver cannot tell which symbols were flipped.

Capacity

Definition 0.4 (Channel capacity)**.** Send one input X through a discrete memoryless channel, and let Y be the output. The joint distribution of (X, Y) is $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\, p(y \mid x)$, so the mutual information $\mathbb{I}(X;Y)$ depends on the channel and on the input probability mass function. The **capacity** of the channel is the largest mutual information any choice of X achieves:

$$C \stackrel{\text{def}}{=} \max_{\mathbb{P}(X=\cdot)} \mathbb{I}(X;Y),$$

where the maximum runs over all probability mass functions on $\mathcal{X}$ (Cover and Thomas 2006, chap. 7). With base-2 logarithms, C is measured in bits per channel use.

Remark 0.2 (What capacity measures). $\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X \mid Y)$ (mutual information as a reduction in entropy) is the uncertainty about the input, minus the uncertainty about the input left after seeing the output. So $\mathbb{I}(X;Y)$ is how much one output tells the receiver about the input, on average, and the capacity is that amount when the sender chooses how often to send each input as well as possible.

Definition 0.5 (Binary entropy function)**.** The **binary entropy function** is

$$h(\epsilon) \stackrel{\text{def}}{=} -\epsilon \log_2 \epsilon - (1-\epsilon) \log_2 (1-\epsilon)$$

for $0 < \epsilon < 1$, with $h(0) = h(1) \stackrel{\text{def}}{=} 0$.

Remark 0.3 (Binary entropy is the entropy of a coin). $h(\epsilon)$ is the entropy, in bits, of a random variable that takes one of two values, one of them with probability ϵ. It is 0 when the outcome is certain and largest, 1 bit, at $\epsilon = 1/2$.

Theorem 0.1 (Capacity of the binary symmetric channel)**.** *The binary symmetric channel with crossover probability ϵ has capacity*

$$C = 1 - h(\epsilon) \text{ bits per channel use,}$$

where h is the binary entropy function. A uniform input achieves it.

Proof

Proof. Fix any probability mass function for X. By mutual information as a reduction in entropy,

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

Given $X = x$, the output equals x with probability $1-\epsilon$ and differs from it with probability ϵ, so the entropy of Y given $X = x$ is $h(\epsilon)$ for both values of x. Averaging over x, the conditional entropy is $\mathbb{H}(Y \mid X) = h(\epsilon)$, whatever X is.
Y takes two values, so $\mathbb{H}(Y) \le \log_2 2 = 1$ bit (Cover and Thomas 2006, sec. 2.6). Therefore

$$\begin{aligned} \mathbb{I}(X;Y) &= \mathbb{H}(Y) - h(\epsilon) \\ &\le 1 - h(\epsilon). \end{aligned}$$

If X is uniform on $\{0,1\}$, then

$$\begin{aligned} \mathbb{P}(Y=1) &= \mathbb{P}(X=0)\,\epsilon + \mathbb{P}(X=1)\,(1-\epsilon) \\ &= \tfrac{1}{2}\epsilon + \tfrac{1}{2}(1-\epsilon) \\ &= \tfrac{1}{2}, \end{aligned}$$

so $\mathbb{H}(Y) = 1$ bit and $\mathbb{I}(X;Y) = 1 - h(\epsilon)$. This value is attained, so it is the maximum C. □

Example 0.1 (A channel that flips one bit in ten)**.** For a binary symmetric channel with $\epsilon = 0.1$,

$$\begin{aligned} h(0.1) &= -0.1 \log_2 0.1 - 0.9 \log_2 0.9 \\ &\approx 0.332 + 0.137 \\ &\approx 0.469 \text{ bits,} \end{aligned}$$

so by Theorem 0.1 the capacity is

$$C \approx 1 - 0.469 = 0.531 \text{ bits per channel use.}$$

Each use carries one bit, but on average only about half a bit of it gets through. At

$\epsilon = 1/2$, $h(1/2) = 1$ and $C = 0$: the output is independent of the input. At $\epsilon = 1$, $C = 1$ again, since the receiver can flip every bit back.

Exercise 0.1 (Capacity of the binary erasure channel)**.** Send X through the binary erasure channel with erasure probability α, where $0 < \alpha < 1$, and write $\pi \stackrel{\text{def}}{=} \mathbb{P}(X = 1)$. Let E be 1 if the output Y is the erasure symbol e, and 0 otherwise. Use base-2 logarithms and the binary entropy function h.

1. Show that $\mathbb{H}(Y \mid X) = h(\alpha)$.
2. E is a function of Y. Use the chain rule for entropy, expanded in both orders, to show that $\mathbb{H}(Y) = \mathbb{H}(E) + \mathbb{H}(Y \mid E)$.
3. Show that $\mathbb{H}(Y) = h(\alpha) + (1 - \alpha)\, h(\pi)$.
4. Conclude that the capacity is $C = 1 - \alpha$ bits per channel use.

Solution

Solution 0.1.

1. Given $X = x$, the output is x with probability $1 - \alpha$ and e with probability α, so $\mathbb{H}(Y \mid X = x) = h(\alpha)$ for both values of x, and averaging over x gives $\mathbb{H}(Y \mid X) = h(\alpha)$.
2. The chain rule gives

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

   Each value y of Y fixes E, so $\mathbb{H}(E \mid Y = y) = 0$ for every y, and $\mathbb{H}(E \mid Y) = 0$. Hence $\mathbb{H}(Y) = \mathbb{H}(E) + \mathbb{H}(Y \mid E)$.
3. The channel erases with probability α whatever the input, so E is independent of X and $\mathbb{H}(E) = h(\alpha)$. Given $E = 1$, $Y =$ e for certain, so $\mathbb{H}(Y \mid E = 1) = 0$. Given $E = 0$, $Y = X$, and since E is independent of X, $\mathbb{P}(Y = 1 \mid E = 0) = \mathbb{P}(X = 1) = \pi$, so $\mathbb{H}(Y \mid E = 0) = h(\pi)$. Therefore

$$\begin{aligned}\mathbb{H}(Y) &= \mathbb{H}(E) + \mathbb{P}(E = 1)\, \mathbb{H}(Y \mid E = 1) + \mathbb{P}(E = 0)\, \mathbb{H}(Y \mid E = 0) \\ &= h(\alpha) + \alpha \cdot 0 + (1 - \alpha)\, h(\pi) \\ &= h(\alpha) + (1 - \alpha)\, h(\pi).\end{aligned}$$

4. Combining parts 1 and 3,

$$\begin{aligned}\mathbb{I}(X;Y) &= \mathbb{H}(Y) - \mathbb{H}(Y \mid X) \\ &= h(\alpha) + (1-\alpha)\, h(\pi) - h(\alpha) \\ &= (1-\alpha)\, h(\pi).\end{aligned}$$

Since $h(\pi) \leq 1$, with equality at $\pi = 1/2$, the maximum over input probability mass functions is $C = 1 - \alpha$. The receiver loses a share α of the bits and learns the rest exactly; Cover and Thomas (2006, sec. 7.1) reach the same capacity.

The noisy-channel coding theorem

Definition 0.6 (Channel code)**.** An (M, n) **channel code** for a discrete memoryless channel sends one of M messages, $w \in \{1, \dots, M\}$, in n uses of the channel. It has

- an *encoder*, which assigns each message w to a *codeword* $(x_1(w), \dots, x_n(w))$ of n symbols sent through the channel, and
- a *decoder*, which sends each list of n outputs to a guess at w.

This is the definition of Cover and Thomas (2006, sec. 7.5).

Definition 0.7 (Rate of a channel code)**.** The **rate** of an (M, n) channel code is

$$R \stackrel{\text{def}}{=} \frac{\log_2 M}{n} \text{ bits per channel use.}$$

This is the definition of Cover and Thomas (2006, sec. 7.5).

Definition 0.8 (Maximal probability of error)**.** The **maximal probability of error** $\lambda^{(n)}$ of an (M, n) channel code is the largest, over messages w, of the probability that the decoder's guess is wrong when w is sent. This is the definition of Cover and Thomas (2006, sec. 7.5).

Example 0.2 (Sending each bit three times)**.** To send one bit over the binary symmetric channel with $\epsilon = 0.1$, send it three times and decode by majority vote. This is a $(2, 3)$ channel code, with rate $R = 1/3$. The decoder is wrong when two or three of the copies are flipped, so the maximal probability of error is

$$\begin{aligned}\lambda^{(3)} &= 3\epsilon^2(1-\epsilon)+\epsilon^3 \\ &= 3(0.01)(0.9)+0.001 \\ &= 0.028,\end{aligned}$$

down from 0.1 for sending the bit once. Repeating more often lowers the error further, but the rate falls toward 0 as it does. MacKay (2003, chap. 1) works through this code.

Theorem 0.2 (Noisy-channel coding theorem)**.** *Let C be the capacity of a discrete memoryless channel.*

1. *For every rate $R < C$ there is a sequence of $(\lceil 2^{nR} \rceil, n)$ channel codes, for $n = 1, 2, ...$, whose maximal probability of error $\lambda^{(n)}$ tends to 0.*
2. *Conversely, any sequence of $(\lceil 2^{nR} \rceil, n)$ codes with $\lambda^{(n)} \to 0$ has $R \leq C$.*

Remark 0.4 (Reliable communication at a positive rate). Example 0.2 buys fewer decoding mistakes with a lower rate, and repetition alone drives the chance of a mistake to 0 only as the rate falls to 0. Theorem 0.2 says that this trade-off is not forced: over the channel of Example 0.1, long enough codes send about 0.53 bits per channel use with a probability of decoding mistakes as small as we like. Above capacity, no code can. Shannon (1948) stated this result and sketched its proof. Cover and Thomas (2006, sec. 7.7) prove part 1 by choosing codewords randomly, and Cover and Thomas (2006, sec. 7.9) prove part 2 with a method due to Fano. MacKay (2003, chap. 9 and 10) gives another account.

References

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

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

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.